Exploration du rôle de l'épissage mineur dans le développement embryonnaire : modèle du syndrome de Taybi-Linder) (TALS)

Audric Cologne

N°d'ordre NNT :
2019LYSE1190

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale ED476
Neurosciences et Cognition

Spécialité de doctorat : Neurosciences
Discipline : Biologie, médecine et santé

Soutenue publiquement le 10/10/2019, par :
Audric David Charles COLOGNE

Exploration du rôle de l'épissage mineur dans le développement embryonnaire : modèle du syndrome de Taybi-Linder (TALS)

Devant le jury composé de :

ATTIE-BITACH, Tania	PUPH	INSERM	Présidente
MARTINS, Alexandra	Chargé de Recherche	INSERM	Rapporteure
RITCHIE, William	Chargé de Recherche	CNRS	Rapporteur
DJEBALI-QUELEN, Sarah	Post-doctorante	INRA	Examinatrice
JULLIARD, Karyn	Professeure des Universités	UCBL1	Examinatrice
MAZOYER, Sylvie	Directeur de Recherche	INSERM	Examinatrice
EDERY, Patrick	PUPH	CRNL	Directeur de thèse
LACROIX, Vincent	Maître de Conférences	UCBL1	Codirecteur de thèse
LEUTENEGGER, Anne-Louise	Chargé de Recherche	INSERM	Invitée

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Résumé

EXPLORATION DU RÔLE DE L'ÉPISSAGE MINEUR DANS LE DÉVELOPPEMENT EMBRYONNAIRE : MODÈLE DU SYNDROME DE TAYBI-LINDER (TALS)

Le Syndrome de Taybi-Linder (TALS) est une maladie génétique rare affectant le développement embryonnaire, caractérisée par un nanisme microcéphalique sévère et un décès précoce des patients. Le gène muté dans ce syndrome est *RNU4ATAC*, qui produit un petit ARN nucléaire (snRNA) non-codant : U4atac. Ce snRNA est l'une des briques composant le spliceosome mineur, une machinerie nucléaire dédiée à l'épissage des introns U12, un groupe d'introns peu étudié car présent dans ~1 % des gènes seulement. Dans le TALS, ces introns sont fréquemment retenus dans les transcrits matures, l'épissage correct des introns U12 semble donc capital pour le développement embryonnaire. L'étude du profil transcriptomique des patients TALS permet ainsi d'établir les conséquences moléculaires d'un dysfonctionnement du spliceosome mineur, nous en apprenant davantage sur les mécanismes d'épissage des introns U12 en condition physiologique ou pathologique, et sur le rôle de l'épissage mineur dans le développement embryonnaire. Cette thèse présente la première analyse approfondie du transcriptome de cellules provenant de patients TALS.

Pour mener cette analyse, nous avons développé un pipeline bioinformatique qui, à partir de données RNA-seq de seconde génération, utilise différentes méthodes dédiées à l'étude différentielle de l'expression des gènes ou de la qualité d'épissage entre patients et contrôles. L'épissage étant particulièrement complexe à analyser à partir de reads courts, deux approches complémentaires ont été utilisées : l'une classique, basée sur l'alignement des reads, et l'autre plus originale, basée sur l'assemblage des reads et permettant de détecter plus d'événements d'épissage non-annotés (KisSplice). Une des conséquences attendue d'un dysfonctionnement du spliceosome mineur est une rétention massive des introns U12 dans les ARN matures. Cependant, la détection et la quantification de rétentions d'intron chez les mammifères constituent encore aujourd'hui un challenge bioinformatique. Nous avons donc utilisé une méthode récente dédiée à l'analyse des rétentions d'introns pour caractériser le plus précisément possible le profil transcriptomique des patients TALS. J'ai ainsi participé au développement de KisSplice et de notre outil d'analyse statistique des différentielles d'épissage, kissDE, et mis en évidence certaines caractéristiques de l'épissage mineur, que ce soit en condition physiologique ou pathologique.

Mot-clefs : *TALS, épissage mineur, introns U12, RNU4ATAC, transcriptomique, bioinformatique*

Abstract

EXPLORATION OF MINOR SPLICING FUNCTION DURING EMBRYONIC DEVELOPMENT WITH THE TAYBI-LINDER SYNDROME (TALS) MODEL

The Taybi-Linder Syndrome (TALS) is a rare genetic disorder of the embryonic development leading to a severe microcephaly, a primordial dwarfism and an early/unexpected death. The mutated gene in this syndrome is *RNU4ATAC*, which is transcribed into a non-coding small nuclear RNA (snRNA) named U4atac, involved in the minor spliceosome. This nuclear machinery is dedicated to the splicing of a small number of particular introns : the U12 introns. Because only about 1 % of the Human's genes display at least one U12 intron, they have not been extensively study and little is known about their function. In TALS patients' cells, most of the U12 introns are retained in mature transcripts ; hence, splicing of U12 introns seems important for the embryonic development. Studying TALS patients' cells transcriptomes both in physiological and pathological conditions should enable us to precisely identify most of the molecular consequences of a minor splicing defect and could shed light on the mechanism linking minor splicing and embryonic development. This thesis is the first work to conduct an in depth analysis of TALS patients' cells transcriptomes.

In order to do a precise analysis, we developed a bioinformatic pipeline that uses multiple methods to detect differentially expressed or spliced genes between patients and controls and from second generation RNA-seq data. Splicing analysis is a very complex task complete with short reads ; hence, we used two complementary approaches. The first one is based on reads alignement to a reference genome, method conventionnally used to work on splicing, and the second one is based on reads assembly (KisSplice), a original method enabling to find more non-annotated splicing events. One of the expected consequences of a minor splicing malfunction is a global U12 introns retention in mature transcripts. However, intron retention detection and quantification in mammals is particulary difficult task in mammals, thus we used a new method dedicated to intron retentions analysis to study the transcriptomic profile of TALS patients. During my thesis, I was one of the developer of KisSplice and kissDE, our differential splicing analysis tool, and I identified important charcteristics of minor splicing either in physiological or pathological conditions.

Keywords *: TALS, minor splicing, U12 introns, RNU4ATAC, transcriptomic, bioinformatic*

Remerciements

Tout d’abord, je remercie l’INRIA et l’INSERM pour avoir financé mes trois années de thèse, ainsi que l’UCBL et le CRNL, organismes dans lesquels se sont effectués mes travaux.

Pour avoir accepté d’évaluer ce manuscrit et d’assister à ma soutenance de thèse, je souhaite remercier chaleureusement les membres de mon jury : Alexandra Martins et William Ritchie (rapporteurs) ; Karyn Julliard, Sarah Djebali-Quelen et Tania Attié-Bitach (examinatrices).

Mes remerciement les plus profonds et mes respects les plus sincères vont à mon directeur de thèse, Patrick Edery, mon co-directeur de thèse, Vincent Lacroix, ma tutrice, Sylvie Mazoyer, et notre collaboratrice, Anne-Louise Leutenegger. Je suis conscient de la chance que j’ai eu de vous côtoyer et je garderai longtemps en mémoire nos discussions et débats, qu’ils aient été scientifiques ou non. Vous m’avez appris à devenir un chercheur responsable, respectueux et éthique, ainsi que, je le pense, une personne meilleure.

Je remercie l’ensemble de mes collègues, à la fois de l’équipe GENDEV : Alicia, Audrey, Marion et Deepak ; et de l’équipe ERABLE, qu’ils soient passés ou présents : Amandine, Blerina, Camille M., Camille S., Carol, Claire, Hélène, Irene, Janice, Laura, Mariana, Marina, Marie-France, Alex, Arnaud, Eric, Leandro, Martin, Nicolas, Ricardo, Taneli. Plus particulièrement, je tiens à exprimer ma reconnaissance envers Clara Benoit-Pilven, qui m’a aidé sans discontinuer et qui a éclairé plus d’une fois ma lanterne au cours de cette thèse ! Je salue avec mélancolie la mémoire de Mattia Gastaldello, et conserve son souvenir précieusement avec moi.

Merci à tous ces rares enseignants passionnés et passionnants qui ont éveillé ma curiosité et mon envie d’apprendre en primaire, collège et lycée.

Pour leur soutien indéfectible, mais aussi pour m’avoir rappelé quotidiennement qu’il n’y a pas que la science dans la vie (même si j’en doute encore!), j’exprime toute ma gratitude envers mes amis proches : Anaïs, Eléa, Julie, Mélanie, Violette, Alex, Coco, Damien, Florian, Germain, Riwann, Roland, Valentin, ainsi que Aurélien et Cédric d’Aurillac, Morgan et Victor de Clermont-Ferrand.

Enfin, je remercie de tout mon cœur ma mère et mon père (qui seront obligés de lire ce manuscrit!), ainsi que Pauline, sans qui tout ce travail n’aurait eu aucun sens à mes yeux.

« La vie, ce concept mystérieux, est ramenée à la présence d'ADN. Il n'y a plus de frontière entre matière animée et inanimée.
Tout n'est qu'une question de degré de complexité. »
Albert JACQUARD (1925-2013), généticien.

Table des matières

Table des Figures

Liste des Tableaux

Liste des abréviations

3'ss : 3' splice-site (site d'épissage 3')

5'ss : 5' splice-site (site d'épissage 5')

A : Adénine

AA : Acide Aminé

ACP : Analyse en Composantes Principales

ADN : Acide DésoxyriboNucléique

ADNc : ADN complémentaire (à un ARN)

ALS : Amyotrophic Lateral Sclerosis (sclérose latérale amyotrophique)

altA/D/AD : alternative Acceptor/Donor/Acceptor and Donor (accepteur/donneur/accepteur et donneur alternatif)

ARN : Acide RiboNucléique

ARNm : ARN messager

ARNnc : ARN non-codant

ARNpolII : ARN polymérase II

ARNpolIII : ARN polymérase III

ARNpre : ARN précurseur

AS : Alternative Splicing (épissage alternatif)

AS-NMD : Alternative Splicing coupled with NMD (épissage alternatif couplé au NMD)

ASTER : projet Algorithmes et outils logiciels pour le Séquençage d'ARN de Troisième génERation

BA : Base Azotée

BMKS : Burn-McKeown Syndrome

BPS : Branch Point Sequence (séquence du site de branchement)

C : Cytosine

CA : Cerebellar Ataxia (ataxie cérébelleuse)

CB : Cajal Bodies

CCMS : CerebroCostoMandibular Syndrome (syndrome cérébrocostomandibulaire)

CRNL : Centre de Recherche en Neurosciences de Lyon

CTD : C-Terminal Domain (domaine C-terminal)

DCM : Dilated CardioMyopathy (Cardiomyopathie dilatée)

DE : Differentially Expressed (différentiellement exprimé)

EJC : Exon Junction Complex (complexe de jonction d'exon)

ES : Exon Skipping (saut d'exon/exon cassette)

FDR : False Discovery Rate (taux de fausses découvertes)

FPKM : Fragments Per Kilobase per Milion mapped reads (fragments par kilobase par million de reads alignés)

G : Guanine

GDB : Graphe de De Bruijn

GEM : GEMini of cajal bodies

GENDEV : GENétique des anomalies du neuroDEVeloppement

Gène U12 : gène contenant au moins un intron U12 (mineur)

Gène U2 : gène ne contenant aucun intron U12

HS : Hypotrichose Simple

IGHD : Isolated Growth Hormone Deficiency (déficience en hormone de croissance isolée)

IGV : Integrative Genomics Viewer

Inria : Institut national de recherche en informatique et en automatique

Inserm : Institut national de le santé et de la recherche médicale

INT : Integrator (complexe moléculaire)

INTD : Integrator-Deficiency

IQR : InterQuartile Range (zone inter-quartiles)

IR : Intron Retention (rétention d'intron)

JBS : JouBert Syndrome

K2RG : KisSplice2RefGenome

kDE : kissDE

KS : KisSplice

LBBE : Laboratoire de Biométrie et Biologie Évolutive

LCL : Lymphoblastoid Cell Line (lignée cellulaire lymphoblastoïde)

LFC : Logarithmic Fold-Change (fold-change logarithmique)

LGMD1G : Limb-Gridle Muscular Dystrophy 1G (dystrophie musculaire des ceintures type 1G)

LSm : Like Sm protein (protéine Sm-like)

LWS : Lowry-Wood Syndrome

MDS : MyeloDysplastic Syndrome (syndrome myélodysplasique)

MFDM : MandibuloFacial Dystosis with Microcephaly (dystose mandibulo-faciale avec microcéphalie)

NAS : NAger Syndrome

NB : Negative Binomial distribution (distribution négative binomiale)

NMD : Non-sense Mediated Decay

ONT : Oxford Nanopore Technologies

ORF : Open Reading Frame (cadre ouvert de lecture)

PacBio : Pacific Bioscience

PC : Principal Components (composantes principales)

PCH7 : PontoCerebellar Hypoplasia type 7 (hypoplasie pontocérébelleuse type 7)

PCR : Polymerase Chain Reaction

PM : Per Milion scaling factor (facteur de mise à l'échelle)

PPT : Poly-Pyrimidine Tract

PSI : Percent Spliced In (pourcentage d'inclusion)

PTC : Premature Terminaison Codon (codon STOP prématuré)

Read : lecture de RNA-seq

RFMN : Roifman Syndrome

RNA-seq : RNA-sequencing (séquençage de l'ARN)

RP : Retinitis Pigmentosa (rétinite pigmentaire)

RPK : Reads Per Kilobase (reads par kilobase)

RPKM : Reads Per Kilobase per Milion mapped reads (reads par kilobase par million de reads alignés)

RPM : Reads Per Milion (reads par million)

Sm site : site de liaison aux protéines Sm

SMA : Spinal Muscular Atrophy (amyotrophie spinale)

SMN : Survival of Motor Neurons (complexe moléculaire)

snRNA : U-rich small nuclear RNA (petit ARN nucléaire riche en Uracile)

snRNP : small nuclear RiboNucleo Protein (petite ribonucléo protéine nucléaire)

T : Thymine

TALS : TAybi-Linder Syndrome

TPM : Transcripts Per Milion (transcrits par million)

U : Uracile

1 Introduction

Au début du XIX^{ème} siècle, les premières études sur la transmission de caractères physiques à la descendance établirent les fondements de la génétique. Cette discipline deviendra une branche à part entière de la biologie, s'intéressant aux unités fondamentales de l'information biologique : les gènes. À la fin de ce même siècle, des liens entre les gènes et l'Acide DésoxyriboNucléique (ADN) mirent en évidence que cette molécule était le support de l'information génétique, devenant ainsi le principal sujet d'étude des généticiens. En 1953, Watson, Crick et Franklin découvrirent la structure en double hélice de l'ADN et peu de temps après, un modèle pour la conservation et l'utilisation de l'ADN fut proposé : la théorie fondamentale de la biologie moléculaire.

Cette théorie établit un lien entre les gènes et les protéines, macromolécules essentielles aux fonctions biologiques, en suivant le schéma suivant (Figure 1).

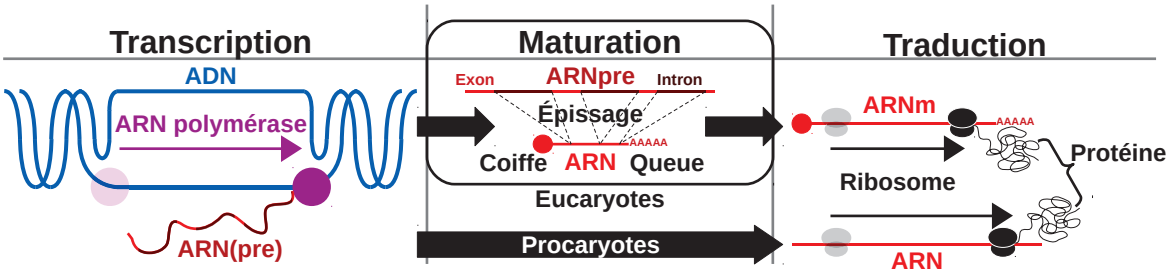

Figure 1 : Théorie fondamentale de la biologie moléculaire.
Les étapes menant de l'ADN à la protéine sont représentées : transcription, maturation (pour les eucaryotes uniquement) et traduction. Pour les procaryotes, toutes ces étapes se réalisent dans le cytoplasme, alors que la transcription se réalise dans le noyau des eucaryotes. L'ARNm (ARN messager) représente, pour les procaryotes, les ARN maturés (ARN) codant une protéine. Les cadres ouverts de lecture (ORF) sont représentés par les flèches symbolisant le parcourt des ribosomes.

L'ADN est une molécule double brin, stable, fortement condensée et localisée dans le noyau de chaque cellule eucaryote, ou dans un compartiment démuni de membrane nucléaire (le nucléoïde) situé dans le cytoplasme des cellules procaryotes. Chez l'Homme, l'ADN est composé de ~3 milliards de Bases Azotées (BA) présentent sous quatre formes : les Adénines (A), les Cytosines (C), les Guanines (G) et les Thymines (T). Elle est aussi répartie en 23 paires de chromosomes, qui ensemble constituent le génome d'un individu. Bien qu'étant le

support de l'information génétique, il est estimé que, chez l'Homme, les ~20 000 gènes codant une protéine représentent moins de 2 % de ce génome. Tous les gènes codant ont la particularité d'être copié de l'ADN vers une autre molécule, l'Acide RiboNucléique (ARN), par l'ARN polymérase II (ARNpolII) chez les eucaryotes ou l'ARN polymérase procaryote, lors d'une étape appelée transcription (Figure 1).

L'ARN, ou transcrit, est une molécule simple brin peu stable, utilisé pour retranscrire l'information génétique correspondant à un gène et produire une protéine. Les Thymines sont remplacées par une 5ème BA : l'Uracile (U). Chez les eucaryotes, le produit brut de la transcription d'un ADN est l'ARN précurseur (ARNpre), qui, après quelques modifications, formera le transcrit mature. Cette maturation se fait par l'ajout d'une coiffe et d'une queue poly-A, mais aussi par le retrait de certains segments de l'ARN (les introns) et la liaison des segments conservés (les exons) lors de l'épissage (Figure 1). Les introns sont reconnus en partie grâce aux séquences caractéristiques de leurs sites d'épissage, délimitant leur début et leur fin. Certains transcrits seront alors exportés dans le cytoplasme, où ils pourront synthétiser ou non une protéine lors de la traduction. Pour la plupart des gènes procaryotes, l'ARN transcrit est directement traduit en protéine (Figure 1).

Les ARN matures codants (ARN messager, ARNm) sont constitués de codons (combinaison de trois BA), qui seront traduits en Acides Aminés (AA) par les ribosomes. Ce sont ces AA qui formeront les protéines. Une correspondance codon-AA est proposée par le code génétique ; on y trouve 4 codons particuliers : AUG (codon START), indiquant le début de la traduction, et UAG/UAA/UGA (codons STOP), indiquant tous trois la fin de la traduction. La séquence située entre un codon de début et un codon de fin de traduction s'appelle un cadre ouvert de lecture (Open Reading Frame, ORF) et détermine ainsi la composition en AA de la protéine codée par l'ARN (Figure 1). Ces protéines vont ensuite pouvoir subir diverses modifications post-traductionnelles, interagir et s'associer avec d'autres protéines ou encore être exportées dans certains compartiments cellulaires pour pouvoir assurer l'ensemble des fonctions moléculaires permettant la vie des cellules.

Chaque étape menant d'un gène à une protéine est finement régulée pour en assurer la qualité. Nous nous intéresserons spécifiquement à l'une de ces régulations, exclusive aux eucaryotes : l'épissage, lors de la modification des ARNpre en transcrits matures.

I. Épissage

L'épissage consiste à exciser les introns de l'ARNpre et lier les exons entre eux. Ce processus est catalysé par une machinerie moléculaire complexe, appelée spliceosome, composée de plus de 200 protéines et de 5 petits ARN nucléaires riches en U (U-rich small nuclear RNA, snRNA) le plus souvent associés à des protéines (snRNP), s'assemblant de manière hautement dynamique (Cvitkovic and Jurica, 2013; Wahl et al., 2009; Will and Luhrmann, 2011). Le terme « intron » sera utilisé ici pour désigner uniquement les introns spliceosomes-dépendants. En effet, trois autres catégories d'introns, présents chez la plupart des eucaryotes et chez certains archées, ont la capacité d'être épissé sans l'intervention du spliceosome (Irimia and Roy, 2014) : (1) Les introns s'auto-épissant du groupe I et (2) II, peu abondants et parfois trouvés dans des génomes bactériens ou de virus mais le plus souvent présents dans les génomes de chloroplastes et de mitochondrie de divers eucaryotes ; (3) Les introns des ARN de transfert nucléaire et d'archées, qui sont rarement trouvés dans des gènes d'archées.

A. Introns

Pourquoi les introns existent-ils ? Ils ne sont pas trouvés dans les génomes des procaryotes, et ne sont donc pas essentiels à la survie de certaines cellules ; ils nécessitent une dépense d'énergie de la cellule pour les épisser et n'encodent donc pas de séquence protéique, ce qui laisse penser que les introns constituent des séquences inutiles, de « l'ADN poubelle ». Pourtant, leur présence dans les génomes de toutes les espèces eucaryotes, à quelques exceptions près, atteste de leur conservation sur de longues distances évolutives (Irimia and Roy, 2014). Quels sont alors les avantages des introns pour une cellule ?

a. Bénéfices

D'un point de vue évolutif, une séquence intronique, hormis ses sites d'épissage, est en principe libre de toute pression de sélection liée à la qualité des protéines, bien qu'elles se situent dans des régions transcrites du génome. Cette caractéristique permet l'accumulation de mutations et d'éléments transposables (notamment les Alu chez l'Homme), ce qui peut conduire à l'émergence de nouveaux exons (Alekseyenko et al., 2007; Lev-Maor, 2003). De même, des ARN non-codants (ARNnc) (Rearick et al., 2011) où des clusters de gènes codants (Kumar, 2009) peuvent coloniser et proliférer dans les introns tout en profitant de l'activité transcriptionnelle de la région. Les introns seraient donc des zones privilégiées pour l'émergence de nouveaux produits transcriptionnels.

Les séquences introniques jouent aussi un rôle plus direct sur l'expression des gènes (Le Hir et al., 2003). En effet, les gènes humains sans introns ont tendance à être faiblement exprimés (Shabalina et al., 2010), un effet pouvant être inversé en leur insérant un intron (Nott, 2003). De plus, les introns humains étant beaucoup plus longs que les exons, leur présence augmente grandement la taille d'un gène et donc, le temps de transcription (de 10 à 20 minutes pour la plupart des gènes humains, jusqu'à 16h pour ceux composés principalement d'introns) (Heyn et al., 2015). À l'inverse, des séquences introniques peuvent aussi stimuler l'initiation et augmenter la vitesse de la transcription (Le Hir et al., 2003). Ainsi, l'architecture d'un gène, en termes de quantité et longueur d'intron, est un élément régulateur de la transcription.

Enfin, le retrait des introns par le spliceosome permet le plus souvent la formation d'un complexe protéique : le complexe de jonction d'exon (Exon Junction Complex, EJC), à environ 20 nucléotides en amont de la jonction exon-exon. L'EJC va stimuler l'export et la traduction des ARNm dans le cytoplasme, mais aussi interagir avec le contrôle qualité des ARNm (Le Hir, 2001). Comme nous le décrirons plus loin dans ce manuscrit, les introns ont aussi l'avantage de pouvoir être épissés de plusieurs manières différentes.

b. Caractéristiques

Les introns sont composés de séquences plus au moins conservées situées à divers sites particuliers et nécessaires à leur reconnaissance (Figure 2) : le site donneur, en 5' de l'intron (5'splice-site, 5'ss) ; le site accepteur, en 3' de l'intron (3'ss) ; le poly-pyrimidine tract (PPT), en amont du site accepteur ; le point de branchement (Branch Point Sequence, BPS), en amont du PPT et contenant une adénosine conservée nécessaire pour l'épissage. Chez l'Homme, deux types d'introns coexistent : les introns majeurs (U2) et les introns mineurs (U12).

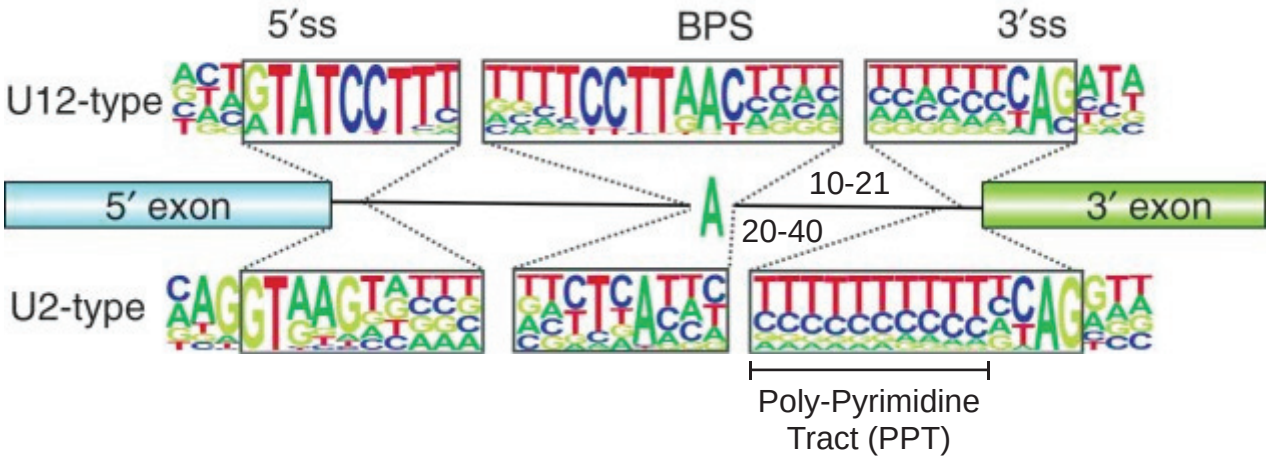

Figure 2 : Séquences consensus des introns U12 et U2.
La conservation d'une base dans la séquence consensus est modélisée par la taille de la lettre correspondante. Les intervalles indiquent le nombre de nucléotide couramment observé entre l'adénosine conservée du BPS et le 3'ss. Source : (Turunen et al., 2013).

Du fait de leur sur-représentation, les introns majeurs correspondent aux introns classiquement décrits dans les publications scientifiques. Les séquences consensus de leurs 5'ss et 3'ss peuvent être fortement variables (ou dégénérées), mais elles commencent et finissent respectivement par les dinucléotides GT et AG dans 99 % des cas (Sheth et al., 2006). Leur BPS, le plus souvent situé entre 20-40 nucléotides du site accepteur (Gao et al., 2008) mais parfois retrouvé à plus de 100 nucléotides du 3'ss, est lui aussi fortement dégénéré (Figure 2).

Les introns U12 étaient d'abord reconnus en se basant uniquement sur les dinucléotides AT et AC, situés en début et fin de l'intron (Jackson, IJ, 1991), bien qu'il fût aussi proposé qu'ils puissent commencer et finir par les dinucléotides classiques, GT et AG (Burge et al., 1998; Dietrich et al., 1997). Ce type d'introns est surtout reconnaissable à la séquence consensus de leur 5'ss et BPS (beaucoup plus conservée et longue que les introns U2), à la distance de 10 à 21 nucléotides séparant le BPS du 3'ss (globalement plus courte comparée aux introns U2) et à l'absence d'un PPT (Figure 2) (Dietrich, 2005; Dietrich et al., 2001). En 2007, ces critères ont permis d'identifier 695 introns U12 dans le génome humain, soit moins de 1 % des introns annotés (Alioto, 2007), dont la majorité (70 %) présentait les dinucléotides donneur GT et accepteur AG.

Malgré leur faible nombre, les introns U12 disposent d'une machinerie d'épissage qui leur est propre, le spliceosome mineur, qui est retrouvé chez le dernier ancêtre commun des eucaryotes (Irimia and Roy, 2014), mais aussi dans la majorité des espèces de plantes et animaux, ainsi que chez certains champignons et insectes. Pourtant, ce spliceosome est absent

dans de nombreuses espèces (Burge et al., 1998; Levine and Durbin, 2001; Russell et al., 2006; Turunen et al., 2013), comme *S. cerevisiae* ou *C. elegans*, dans lesquelles les introns U12 sont aussi manquants (Figure 3) (Bartschat and Samuelsson, 2010). Ceci pourrait être dû à une conversion des introns U12 en introns U2 (Burge et al., 1998; Turunen et al., 2013). Néanmoins, les gènes U12 (gènes contenant au moins un intron U12) et la position des introns U12 dans ces gènes sont globalement conservés au cours de l'évolution (Basu et al., 2008), et ce de manière plus remarquable chez les vertébrés (Lin et al., 2010).

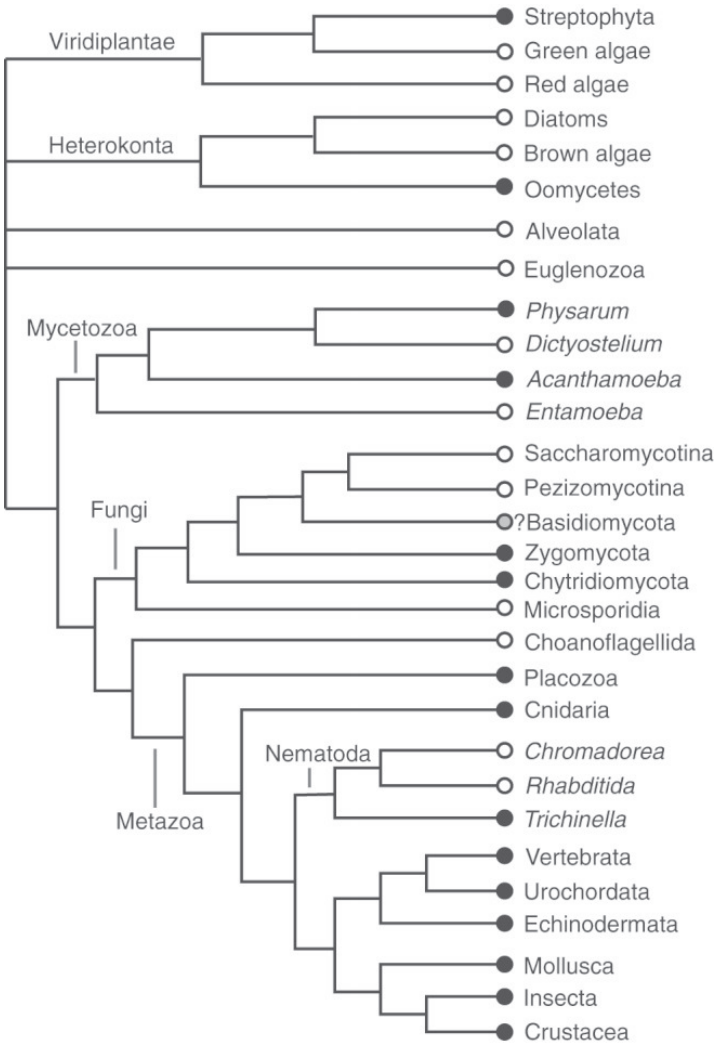

Figure 3 : Distribution phylogénique de l'épissage mineur.
Les cercles pleins et vides représentent les taxons pour lesquels des introns/composants du spliceosome mineur ont été détectés ou non, respectivement. Source : (Turunen et al., 2013).

Les introns U12 de l'Homme sont particulièrement représentés dans des gènes impliqués dans la réplication et réparation de l'ADN, l'organisation du cytosquelette, le transport vésiculaire

et les canaux ioniques voltage-dépendants (Burge et al., 1998; Wu, Quiang and Krainer, 1999; Yeo et al., 2007). Les introns U12 pourraient permettre une régulation plus fine de ces processus biologiques. En effet, des expériences *in vitro* et *in vivo* ont révélé une réduction du taux d'épissage des introns mineurs comparé aux introns majeurs (Frilander and Steitz, 1999; Pessa, 2006; Singh and Padgett, 2009; Tarn and Steitz, 1996) réduisant ainsi le niveau d'ARNm, un effet pouvant être aboli en transformant l'intron U12 en intron U2 (Patel, 2002). La cellule pourrait ainsi réguler l'expression des gènes U12 en modulant l'efficacité d'épissage des introns U12.

Malgré ces différences entre introns mineurs et majeurs, le fonctionnement des spliceosomes qui leurs sont associés est très semblable.

B. Spliceosomes

Chez l'Homme, il existe deux types de spliceosomes : le spliceosome majeur et mineur, qui épissent les introns U2 et U12, respectivement. Ces spliceosomes sont composés des mêmes protéines à 95 % et du même snRNA U5. Les quatre autres snRNA composant le spliceosome majeur (U1, U2, U4, U6) ont tous un homologue mineur (U11, U12, U4atac, U6atac). Les séquences des snRNA homologues sont divergentes (49, 46, 51 et 50 % d'homologie, respectivement) mais ils partagent une structure secondaire similaire (Tarn and Steitz, 1996). De plus, le processus de maturation suivi par tous les snRNA est aussi le même.

a. Biogenèse des snRNA et snRNP

La formation des snRNP peut être résumée en quatre étapes : 1) Transcription du gène codant pour un snRNA ; 2) Maturation co-transcriptionnelle ; 3) Export et maturations cytoplasmiques ; 4) Import et maturation nucléaire (Figure 4) (Gruss et al., 2017).

1) Le génome humain est composé de très nombreuses copies de gènes codants les snRNA majeurs (de ~70 à ~1300 copies), contrairement aux gènes codants les snRNA mineurs (de 1 à ~40 copies) (Vazquez-Arango and O'Reilly, 2018). Chaque snRNA dispose d'un gène canonique exprimé de manière constitutive, mais leurs variants semblent être différentiellement régulés en fonction du stade de développement et du tissu, offrant un nouveau niveau de régulation à la cellule (Lu and Matera, 2015; O'Reilly et al., 2013). Deux domaines cis-régulateurs particulièrement importants pour la maturation des snRNA sont trouvés sur leurs gènes : la 3'box et le site de liaison aux protéines Sm (Sm site) ou protéines like Sm (LSm) pour U6 et U6atac. Les snRNA U6 et U6atac sont les seuls à être transcrits par

l'ARN polymérase III (ARNpolIII), les autres snRNA étant transcrits par l'ARNpolII (Gruss et al., 2017), dont le domaine C-terminal (C-Terminal Domain, CTD) interagit de manière stable avec un complexe d'environ 12 protéines : Integrator (INT) (Baillat et al., 2005).

2) Sur l'extrémité 3' du snRNA en cours de transcription, INT va reconnaître la 3'box et couper l'ARN jusqu'à une vingtaine de nucléotides en amont (Baillat et al., 2005). Un premier type de coiffe sera également co-transcriptionnellement ajouté à l'extrémité 5' du snRNA, ce qui permettra son export cytoplasmique. Les snRNA U6 et U6atac, reçoivent un deuxième type de coiffe et resteront dans le noyau, passant directement à l'étape 4) (Gruss et al., 2017) après recrutement, non décrit dans la littérature, du LSm ring via leur LSm site.

3) Une fois dans le cytoplasme, les snRNA transcrits par l'ARNpolII vont subir trois maturations : une partie de l'extrémité 3' est dégradée par une ou plusieurs exoribonucléase(s) (processus peu connu nommé 3'-trimming) (Neuenkirchen et al., 2008) ; le Sm site va induire la formation du Sm ring et l'arrivée du complexe SMN (Survival of Motor Neurons) ; leur coiffe va être modifiée en un troisième type ce qui permettra l'import du snRNP dans le noyau (Gruss et al., 2017).

4) Les snRNA sont importés dans des compartiments nucléaires appelés Cajal Bodies (CB) où ils subiront les dernières étapes de maturation permettant d'aboutir au mono/di/tri-snRNP fonctionnels, qui seront stockés dans les « nuclear speckles » (Gruss et al., 2017).

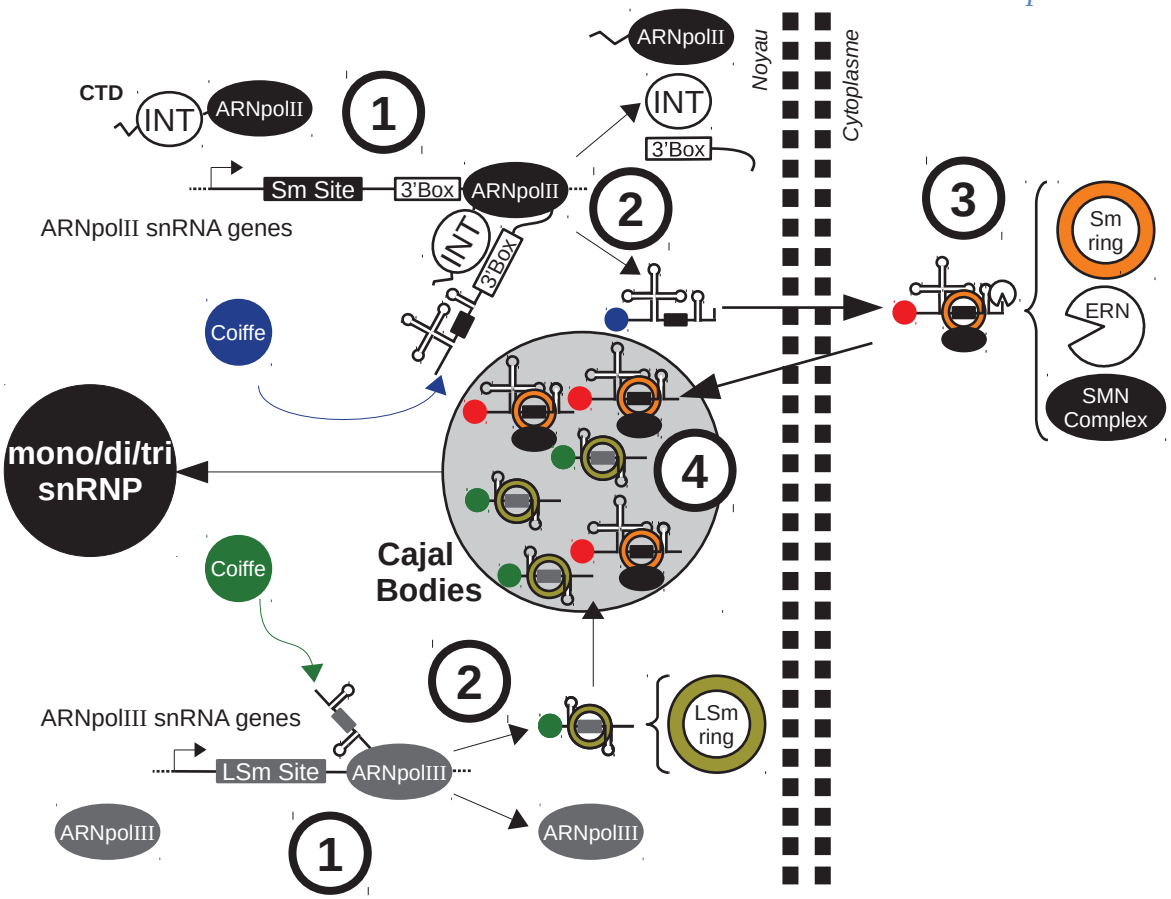

Figure 4 : Vue d'ensemble de la biogenèse des snRNA.
Les 4 grandes étapes, identifiées par des numéros, sont décrites dans le texte. Les coiffes de type 1, 2 et 3 sont colorées en bleu, vert et rouge, respectivement. CTD : C-Terminal Domain, INT : Integrator, ARNpol : ARN polymérase, ERN : ExoRiboNucléase. *Inspiré de (Gruss et al., 2017; Neuenkirchen et al., 2008).*

b. Assemblage des spliceosomes

La composition protéique des spliceosomes est quasiment identique (Schneider et al., 2002), il est donc fortement probable qu'ils s'assemblent de manière analogue (Turunen et al., 2013). L'assemblage du spliceosome majeur a été largement étudié *in vitro*, ces étapes sont aujourd'hui bien caractérisées et passent par la formation de quatre complexes (Figure 5) : E, A, B* et C (Will and Luhrmann, 2011).

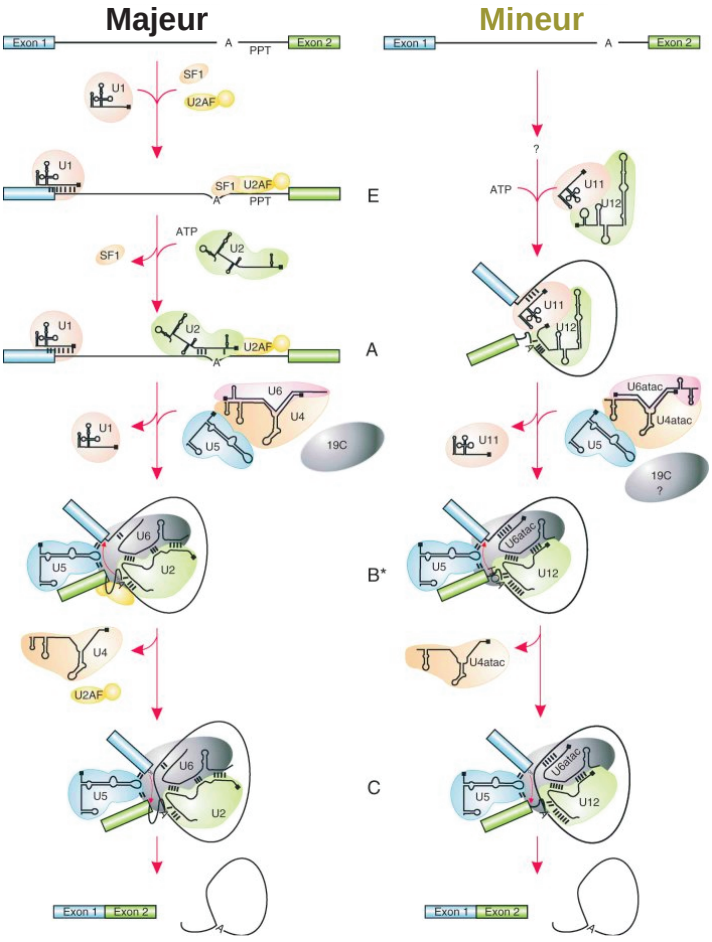

Figure 5 : Assemblage du spliceosome majeur et mineur.
Les lettres E, A, B* et C font références aux différents complexes formés lors de l'assemblage. *Adapté de (Turunen et al., 2013).*

Brièvement, les complexes E et A permettent la reconnaissance des sites d'épissage de l'intron : les snRNP U1 puis U2 vont s'associer aux sites 5' et au BPS de l'intron majeur, tandis que le di-snRNP U11/U12 va reconnaître les sites 5' et le BPS de l'intron mineur de manière coordonnée. Comme les sites d'épissage peuvent être fortement dégénérés, cette reconnaissance se fait à l'aide de facteurs d'épissages associés aux snRNA (e.g. *SF1* et *U2AF*), qui viennent se fixer sur des séquences cis-régulatrices pouvant se situer dans les introns et/ou les exons. D'autres facteurs d'épissage, extérieurs au spliceosome, peuvent aussi interagir avec des séquences présentes dans les exons et/ou introns des gènes pour faciliter la reconnaissance des sites d'épissage. La principale différence entre spliceosome majeur et mineur se situe lors de cette étape, où l'ensemble des 7 protéines spécifiques au spliceosome mineur sont utilisées (Schneider et al., 2002). Le spliceosome mineur est aussi moins dépendant des facteurs d'épissages externes que son homologue majeur (Brock et al., 2008),

privilégiant des interactions ARN/ARN pour reconnaître les introns mineurs, dont les séquences consensus des sites d'épissage sont fortement conservées. Une fois les bornes de l'intron reconnues, le tri-snRNP U4/U6.U5 ou U4atac/U6atac.U5 intègre le pre-spliceosome majeur ou mineur, respectivement. Des réarrangements de la structure des spliceosomes mènent au relargage des snRNA U1 et U4 (U11 et U4atac) et à une nouvelle interaction entre U2 et U6 (U12 et U6atac) ce qui permettra la formation d'un spliceosome catalytiquement actif (complexe B*). Le snRNA U4 (U4atac) n'est ainsi pas directement impliqué dans l'action catalytique du spliceosome, mais agit comme un régulateur de l'acteur principal de cette activation, U6 (U6atac), et donc comme un régulateur de l'efficacité de l'épissage. La première réaction d'épissage est alors réalisée, faisant passer les spliceosomes dans le complexe C, qui catalysera la seconde réaction d'épissage.

L'assemblage du spliceosome se réalise pour chaque intron et principalement de manière co-transcriptionnelle, que ce soit pour l'épissage majeur (Khodor et al., 2011) ou mineur (Singh and Padgett, 2009). Une étude a pourtant rapporté que les snRNA mineurs étaient principalement trouvés dans le cytoplasme, ou une partie de l'épissage mineur pourrait avoir lieu (König et al., 2007) ; cependant ces résultats n'ont pas été reproduits dans des expériences postérieures (Friend et al., 2008; Pessa et al., 2008; Singh and Padgett, 2009), expériences toutefois critiquées par König et Müeller (Konig and Muller, 2008). Des doutes subsistent donc sur l'épissage cytoplasmique d'introns U12. L'épissage des introns U12 a aussi été rapporté comme plus lent ou moins efficace comparé aux U2 (Patel, 2002; Pessa, 2006), ce qui suppose une différence de cinétique entre l'assemblage des deux spliceosomes.

C. Épissage alternatif

Si plusieurs sites d'épissage sont sélectionnables par les spliceosomes, plusieurs introns peuvent être épissés en fonction des sites sélectionnés : on parle alors d'épissage alternatif (Alternative Splicing, AS). Ce phénomène était d'abord réputé rare lors de sa découverte en 1977 (Berget et al., 1977) mais ces 10 dernières années, il a été mis en évidence que plus de 95 % des gènes humains contenant au moins 1 intron pouvaient être épissés de différentes manières (Barash et al., 2010; Pan et al., 2008). L'AS serait particulièrement commun chez les primates (Barbosa-Morais et al., 2012). Il est donc possible que certains exons, ou que certaines parties d'un exon, soient parfois reconnus comme des introns par le spliceosome, donnant naissance à plusieurs transcrits matures (appelés transcrits alternatifs) à partir d'un même ARNpre.

a. Événements d'épissage alternatif

Un événement d'épissage alternatif permet de décrire de quelles manières peuvent être épissés les introns d'un gène en observant ces transcrits alternatifs. Un événement d'épissage se définit par une partie constitutive (invariable) et une partie alternative (variable). La partie constitutive est composée de deux exons retrouvés dans tous les transcrits alternatifs comparés. Dans les cas les plus simples, il n'existe qu'une partie alternative et deux formes d'épissage sont alors observées : l'inclusion (la plus grande, comportant la partie variable et constitutive) et l'exclusion (la plus petite, comportant uniquement la partie constitutive). On peut distinguer trois grands types d'AS (Figure 6) : 1) Le saut d'exon/exon cassette (Exon Skipping, ES), qui inclut/exclut un ou plusieurs exon(s) entier(s) situé(s) entre les exons constitutifs ; 2) Les accepteurs, donneurs et accepteurs+donneurs alternatifs (altA, altD, altAD respectivement), qui incluent/excluent un bout d'un seul ou des deux exons constitutifs ; 3) La rétention d'intron (Intron Retention, IR), qui inclut/exclut l'intron situé entre les deux exons constitutifs (bien que cet événement correspond plus à un non-épissage qu'un épissage alternatif, il sera tout de même considéré comme un AS dans ce manuscrit). D'autres catégories sont parfois utilisées pour décrire les premiers et derniers exons alternatifs, mais ces événements correspondent plus à de la transcription alternative qu'à de l'épissage alternatif et ne seront en conséquence pas discutés ici. Les ES et altAD, ou des combinaisons d'ES et de altA/D/AD, peuvent aussi être « mutuellement exclusifs » dans le cas où chaque transcrit alternatif contient une partie alternative qui lui est propre (il existe alors deux parties variables) (Figure 6). Enfin, si plus de deux formes d'épissages sont possibles, c'est-à-dire si l'intron observé peut être épissé de plus de deux manières différentes, les événements d'épissage sont dénommés complexes.

Un événement d'épissage alternatif permet de différencier des transcrits de manière locale, c'est-à-dire uniquement sur une seule partie alternative située entre deux exons présents chez tous les transcrits comparés. Ainsi, si plusieurs événements d'épissage sont trouvés sur un même gène, il ne sera pas possible de déterminer avec exactitude la liste des transcrits alternatifs de ce gène. Par exemple, dans la Figure 7, quatre transcrits alternatifs peuvent être trouvés à partir du graphe d'épissage, alors que seulement deux de ces transcrits ont servi à construire le graphe. Donc, si à chaque transcrit correspond un chemin dans le graphe, chaque chemin ne représente pas forcément un transcrit observé.

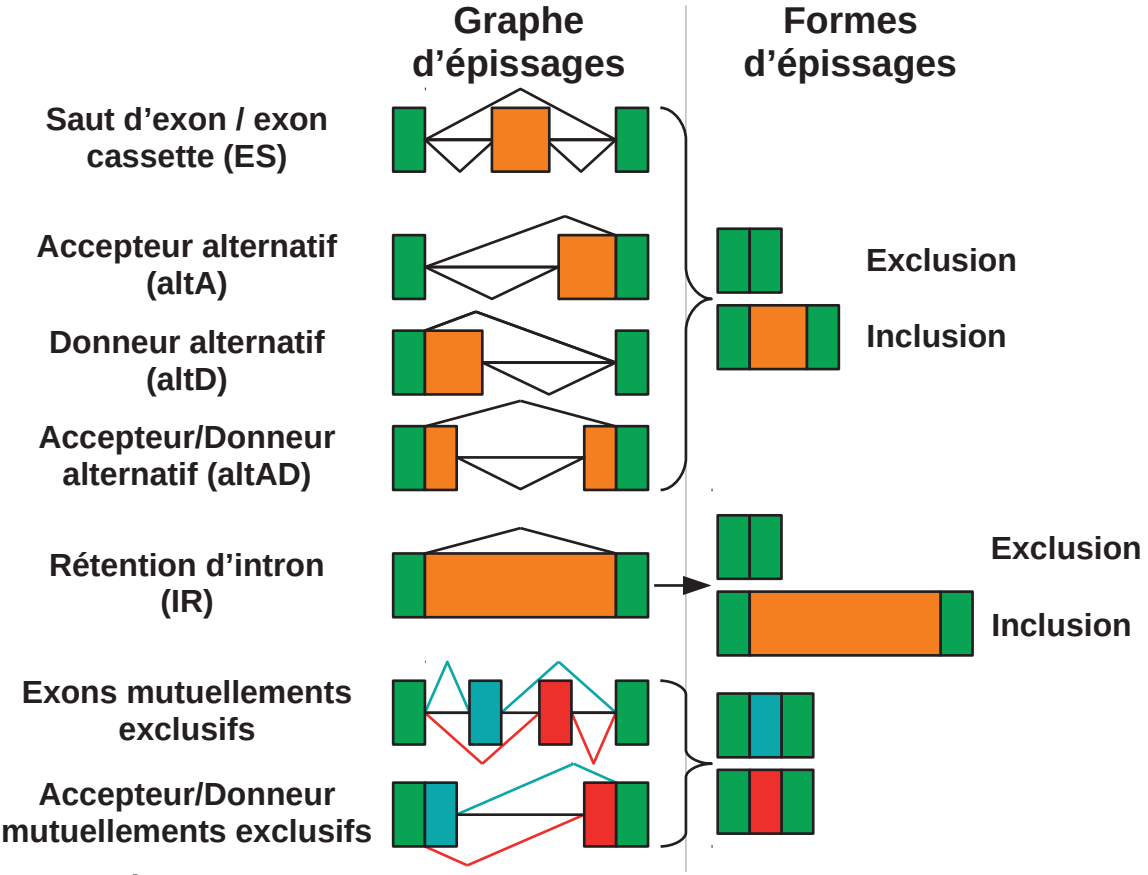

Figure 6 : Événements d'épissages alternatifs simple.
Les traits horizontaux pleins et obliques représentent les introns épissés de manière constitutive et les jonctions exon-exon, respectivement. Les rectangles verts et oranges/bleus/rouges représentent les exons constitutifs et les parties alternatives, respectivement.

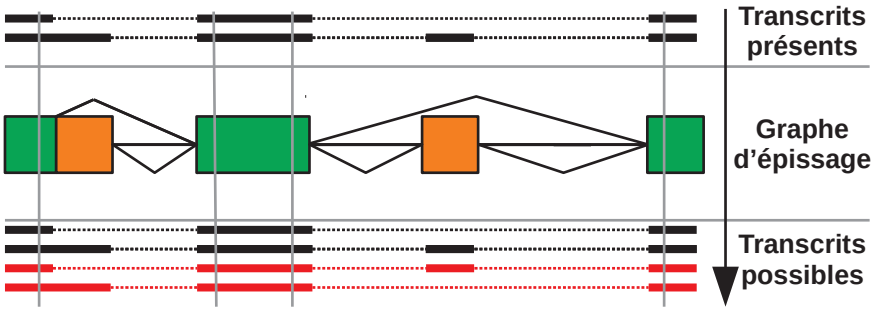

Figure 7 : Reconstruction de transcrits alternatifs à partir d'un graphe d'épissage.
Représentation des transcrits produits par un gène (haut), du graphe d'épissage construit à partir de ces transcrits (milieu) et de la liste des transcrits obtenus en suivant tous les chemins possibles de ce graphe (bas). Les transcrits dérivés de ce graphe mais absents du groupe de transcrits qui a servi à le construire sont représentés en rouge. Les exons et introns sont représentés par des rectangles et des traits horizontaux (en pointillé dans les transcrits), respectivement.

b. Avantages de l'épissage alternatif

L'épissage alternatif est principalement reconnu comme étant une source principale de la diversité des transcrits (Pickrell et al., 2010; Wan and Larson, 2018). C'est l'un des paramètres permettant, à partir d'un même génome, de produire divers transcriptomes ou protéomes (ensemble des ARN matures ou des protéines d'une cellule à un instant, respectivement) en fonction, par exemple, du tissu (Nilsen and Graveley, 2010; Tapial et al., 2017).

En effet, pour un même gène codant produisant des transcrits alternatifs, chacun des transcrits peut potentiellement créer une protéine unique, en particulier si le cadre de lecture de l'ORF est conservé (Zhang et al., 2007). Ces protéines seront alors des isoformes, chacune pouvant remplir une fonction différente, et même opposée, à celle d'autres isoformes. Pourtant, l'abondance des transcrits alternatifs ne correspond pas à l'abondance des isoformes, et il est difficile, avec les méthodes d'étude des protéines actuelles, de déterminer quelle proportion de ces transcrits aboutit à la production d'une protéine fonctionnelle (Wan and Larson, 2018). Tous les transcrits alternatifs ne sont donc pas traduits.

Ces transcrits peuvent tout de même agir sur la quantité de protéine produite par leur gène, en interagissant avec le contrôle qualité cytoplasmique des cellules, et plus spécifiquement avec le Non-Sense Mediated Decay (NMD) via le mécanisme d'AS-NMD (Alternative Splicing coupled with NMD, alternativement nommé RUST pour Regulated Unproductive Splicing and Translation) (Lareau et al., 2007; Lewis et al., 2003). Lors de la production d'un transcrit alternatif, il est possible qu'un nouveau codon STOP soit inséré dans l'ORF, produisant un codon STOP prématuré (Premature Terminaition Codon, PTC) dans la majorité des cas, qui sera une caractéristique des transcrits aberrants. Ce PTC provoquera l'interaction entre le ribosome et l'EJC situé en aval, ce qui recrutera le NMD et mènera à la dégradation du transcrit, l'empêchant ainsi de produire une protéine (Figure 8) (Weischenfeldt et al., 2012). La production de transcrits alternatifs est donc un moyen de contrôler en quantité et/ou qualité l'expression des gènes en fonction du temps et/ou des tissus. Les rétentions d'intron (IR) sont particulièrement sujettes au NMD (Wong et al., 2016) puisque, les introns humains étant longs, on trouve généralement plusieurs PTC dans n'importe quel cadre de lecture d'une séquence intronique.

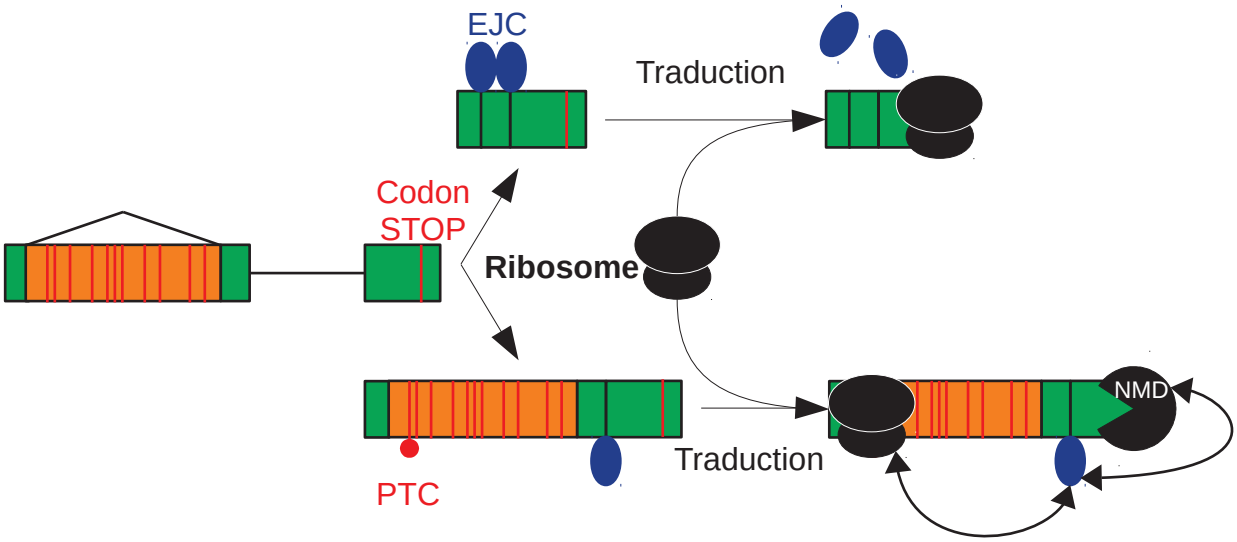

Figure 8 : Mécanisme déclenchant le Nonsense-Mediated Decay.
Lors de la traduction, le ribosome retire les EJC sur son passage. Si le ribosome rencontre un PTC, il interagira avec l'EJC en aval, déclenchant le recrutement du NMD et la dégradation du transcrit. EJC : Exon Junction Complexe, PTC : Premature Termination Codon, NMD : Nonsense-Mediated Decay.

c. Rétention d'intron

Les IR ont longtemps été négligées par la communauté scientifique puisqu'elles étaient considérées comme rares et qu'elles ne contribuaient vraisemblablement pas à la diversité des ARN et des protéines, bien que l'IR soit l'événement d'épissage le plus fréquent chez les plantes, où il joue un rôle important lors du développement et de la réponse au stress (Wong et al., 2016). Ce désintérêt était aussi nourri par la difficulté liée à la détection des IR dans des génomes de mammifères, composés d'introns de plusieurs kilobases, comparé aux plantes, où les introns ont une taille moyenne de 167 paires de bases (Wong et al., 2016). Néanmoins, les IR ont récemment suscité un regain d'intérêt notable (Vanichkina et al., 2018) suite à des études montrant que 3/4 des gènes multi-exoniques sont affectés par des IR (Braunschweig et al., 2014; Middleton et al., 2017). Ces rétentions sont souvent stade-de-développement- (Braunschweig et al., 2014; Jacob and Smith, 2017) et tissu-spécifique, les cellules immunitaires et neurales étant les plus affectées, alors que les cellules souches embryonnaires sont les moins affectées (Vanichkina et al., 2018). La rétention d'un intron dépend de plusieurs paramètres, comme la force des sites d'épissage, la taille ou le GC % de l'intron (Braunschweig et al., 2014; Ge and Porse, 2014). Les IR peuvent être utilisées par le système d'AS-NMD (Ge and Porse, 2014; Wong et al., 2013) : par exemple, l'IR-NMD régule l'expression de gènes contrôlant la différentiation de granulocytes (Figure 9). Les IR peuvent aussi être un moyen de retenir les transcrits contenant un intron dans le noyau pour l'épisser

plus tard, lorsque cela sera nécessaire pour la cellule (mécanisme de detained introns) (Boutz et al., 2015; Jacob and Smith, 2017; Mauger et al., 2016). Enfin, une IR peu aboutir à la production d'un nouvel isoforme, le transcrit contenant l'intron échappant donc aux mécanismes de contrôle qualité, comme le reportent de rares exemples (Wong et al., 2016). Notamment, si le dernier intron est retenu, il n'y aura pas d'EJC en aval, empêchant ainsi le recrutement du NMD.

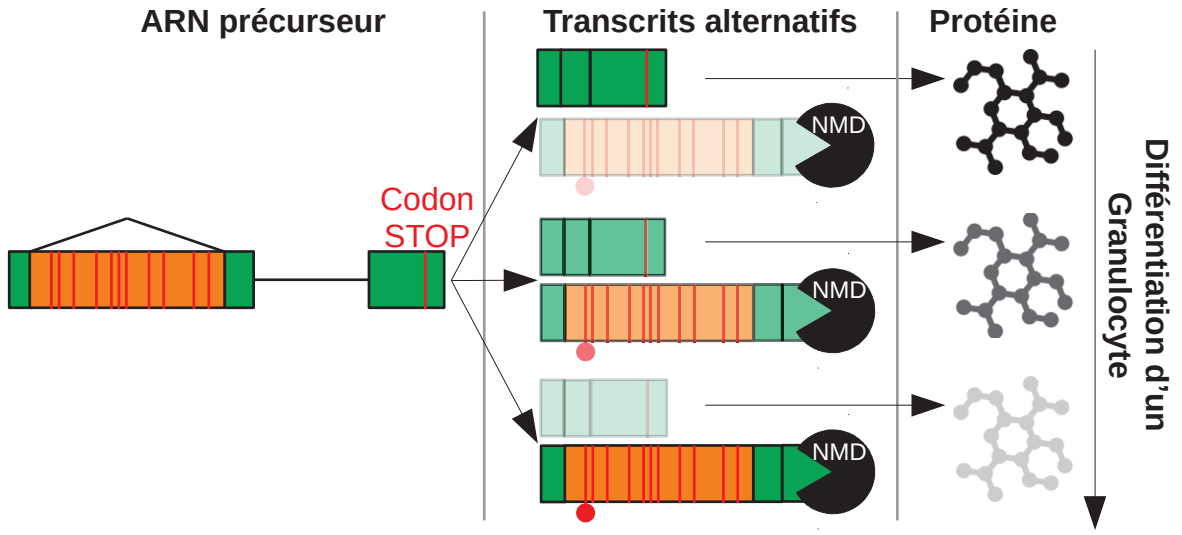

Figure 9 : Mécanisme d'AS-NMD lors de la différentiation d'un granulocyte.
Plus un transcrit alternatif ou une protéine est abondante, moins il apparaît transparent. Un point rouge indique un PTC. PTC : codon STOP prématuré, reconnu par le NMD (Nonsense-Mediated Decay). *Inspiré de (Wong et al., 2013).*

Bien que les introns U12 subissent rarement un AS (Turunen et al., 2013), des mécanismes d'épissage alternatif impliquant un intron majeur et un intron mineur ont été mis en évidence. Chez la drosophile, des gènes contenant le chevauchement d'un intron U2 avec un intron U12, un phénomène nommé twintron (Lin et al., 2010; Scamborova et al., 2004), démontrent une compétition entre les spliceosomes mineur et majeur, l'épissage de l'un ou l'autre des introns engendrant l'inclusion ou l'exclusion de 29 acides aminés dans la protéine (Scamborova et al., 2004). Chez l'Homme, un intron hybride de *MAPK9* dispose d'un donneur de type U12 et d'un accepteur de type U2 (Figure 10), aucun des deux spliceosomes ne peut donc l'épisser. Cependant, cet intron va participer à l'inclusion/exclusion d'exons mutuellement exclusifs de même taille (72 nucléotides) et sans codon STOP en phase, permettant ainsi de changer certains acides aminés dans la protéine finale. Dans les neurones, des facteurs d'épissage vont privilégier la création de la forme d'épissage normalement la moins produite, inversant ainsi l'abondance relative des deux formes (Chang et al., 2007).

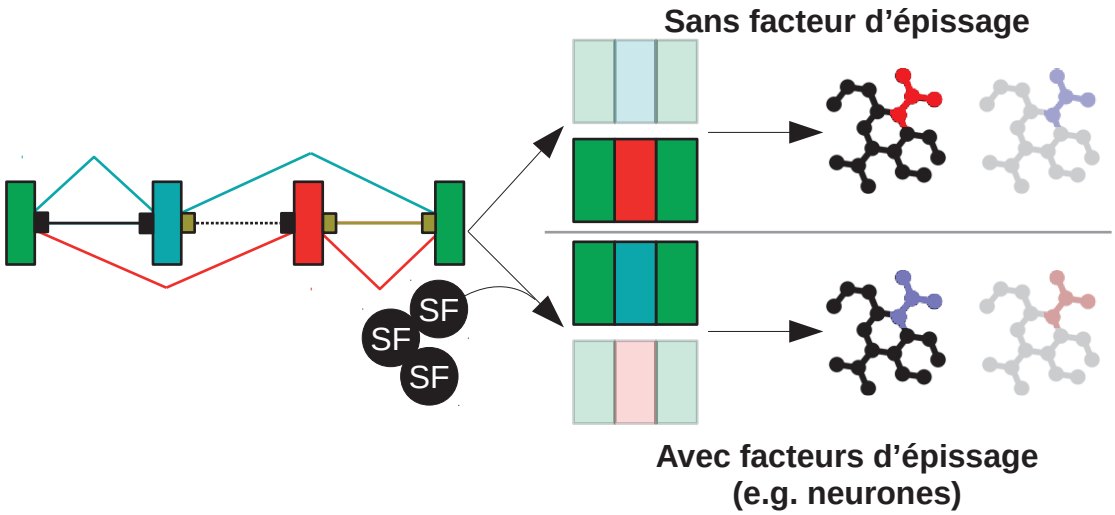

Figure 10 : Exemple d'exons mutuellement exclusifs dépendants de l'épissage d'introns U2 et U12 dans le gène *MAPK9*.
Les sites d'épissage/introns U2 et U12 sont représentés en noir et jaune, respectivement. L'intron hybride non reconnu par les spliceosomes est représenté en pointillé. SF : facteur d'épissage. *Inspiré de (Chang et al., 2007).*

Le fait que les introns U12 soient fortement conservés au cours de l'évolution et qu'une machinerie d'épissage soit dédiée à leur suppression indique que les introns mineurs assurent des fonctions biologiques importantes, comme la différentiation neuronale (Verbeeren et al., 2017). De plus, une supression de composants spécifiques au spliceosome mineur, entraînant une rétention massive des introns U12, conduit à des anomalies du développement chez le poisson zèbre et la drosophile, bien que cette dernière ne contienne que 19 introns U12 (Turunen et al., 2013). Enfin, l'implication de mutations de snRNA U12 dans diverses maladies humaines est une autre preuve du rôle majeur joué par l'épissage mineur.

II. Scpliceosomopathies

Plusieurs types de mutation peuvent affecter l'épissage. Les plus courantes sont celles apparaissant sur des éléments *cis*-régulateurs, comme un site d'épissage ou une séquence régulatrice de l'épissage, représentant ~10 % et ~40 % des maladies humaines liées à une mutation, respectivement. Pour le gène impliqué, ces mutations peuvent avoir plusieurs conséquences : modifier l'abondance respective des transcrits alternatifs, provoquer l'utilisation des sites d'épissage existant mais non-reconnus en condition physiologique (sites cryptiques) ou provoquer l'utilisation de sites d'épissage non-annotés (sites *de novo*) (Anna and Monika, 2018). D'autres mutations, plus rares, peuvent avoir lieu dans des gènes codant des protéines ou snRNA essentiels aux spliceosomes, menant à leur dysfonctionnement et provoquant des maladies regroupées sous le terme de spliceosomopathies, dont une liste est disponible dans le Tableau I (Jutzi et al., 2018; Lardelli et al., 2017; Oegema et al., 2017; Scotti and Swanson, 2016; Vazquez-Arango and O'Reilly, 2018; Verma et al., 2018). Ce n'est alors plus l'épissage d'un, mais de tous les gènes de toutes les cellules de l'organisme qui peut potentiellement être affecté par ces mutations. Paradoxalement, malgré l'apparente ampleur globale de ces maladies, les spliceosomopathies sont principalement caractérisées par des anomalies du développement et des défauts tissu-spécifiques, le cerveau étant l'un des organes les plus affectés.

Les spliceosomes majeur et mineur étant extrêmement semblables, il est donc fort probable que la majorité des spliceosomopathies affectant l'épissage U2 affecte aussi l'épissage U12. Cependant, dans la plupart des cas, seul l'épissage U2 est étudié. Les gènes mutés codent parfois une protéine préférentiellement utilisée par un seul spliceosome, ou un gène de snRNA spécifique, menant vraisemblablement au dysfonctionnement d'une seule des deux machineries d'épissage. Dans le cadre de cette thèse, nous nous intéresserons particulièrement aux spliceosomopathies pour lesquelles des défauts de l'épissage mineur ont été identifiés.

Nous proposons de regrouper les spliceosomopathies en deux catégories, elles-mêmes divisées en deux sous-catégories. Le premier groupe correspond aux mutations sur des gènes codant des protéines associées aux spliceosomes, qui peuvent affecter les facteurs d'épissage externes (n'entrant pas dans la composition des spliceosomes mais pouvant contribuer à l'épissage) ou internes aux spliceosomes. Le deuxième groupe correspond à des défauts des snRNA associés aux spliceosomes, qui peuvent être dus à des mutations sur des gènes impliqués dans la maturation des snRNA ou à des mutations sur les gènes codant les snRNA.

Tableau I : Liste des spliceosomopathies.
Les maladies et gènes mutés impliquant un dysfonctionnement de l'épissage mineur uniquement sont colorés en jaune.

		Maladie	Acronyme	Signes cliniques	Gène muté	Fonction
Mutations affectant les protéines associées aux Spliceosomes	Facteurs d'épissages	Dystrophie musculaire des ceintures Type 1G	LGMD1G	Faiblesse des muscles des ceintures Limitation de la flexion des doigts	*HNRPDL*	Facteur d'épissage
		Cardiomyopathie dilatée	DCM	Insuffisance cardiaque	*RBM20*	Facteur d'épissage
		Sclérose Latérale Amyotrophique	ALS	Paralysies/faiblesses musculaires	*TARDBP* *FUS*	Facteurs d'épissage
	Protéines des snRNP	Dysostose mandibulo-faciale Et microcéphalie	MFDM	Malformations crano-faciales sévères Retard du développement Microcéphalie, déficit intelectuel	*EFTUD2*	Composant du snRNP U5
		Syndrome de Burn-McKeown	BMKS	Malformations crano-faciales sévères Défauts cardiaques/rénales Surdité	*TXNL4A*	Composant du snRNP U5
		Syndrome de Nager	NAS	Malformations mandibulo-faciales Malformations des membres	*SF3B4*	Composant du snRNP U2 et U11/U12
		Syndrome myélodysplastique	MDS	Cancer du sang Nombreux globules blanc immatures	*U2AF1*	Composant du snRNP U2
					SF3B1	Composant du snRNP U2
					ZRSR2	et U11/U12
		Rétinite pigmentaire	RP	Réduction du champ visuel Troubles de la vision Cécité	*SNRNP200* *PRPF2* *PRPF8*	Composant du snRNP U5
					PRPF4 *PRPF6*	Composant du snRNP U4/U6 (U4atac/U6atac?)
					PRPF3 *PRPF31*	Composant du snRNP U4/U6 et U4atac/U6atac
		Déficience en hormone De croissance isolée	IGHD	Insuffisance hypophisaire Retard de croissance Microcephalie légère	*RNPC3*	Composant du snRNP U11/U12
Mutations provoquant un défauts des snRNA	Protéines impliquées dans la maturation des snRNA	Hypoplasie pontocérébelleuse type 7	PCH7	Dégénérescence neurologique Faiblesses musculaires Anomalies de respiration Retard du développement cérébrale	*TEO1*	Maturation des pre-snRNA (coupure de l'extremité 3')
		Integrator-deficiency	INTD	Déficit intellectuel sévère Épilepsie Malformations faciale et des membres	*INTS1* *INTS8*	Composants du complexe Integrator
		Syndrome cérébrocostomandibulaire	CCMS	Anomalies des côtes Anomalies oro-faciales	*SNRNPB*	Protéine Sm
		Hypotrichose simple	HS	Perte de poils/cheveux/sourcils/cils	*SNRPE*	Protéine Sm
		Amyotrophie spinale	SMA	Faiblesse/atrophie des muscles	*SMN1*	Protéine centrale du complexe SMN
	Gènes codant les snRNA	Ataxie cérébelleuse	CA	Incoordination des muscles Dégénérescence cérébelleuse	*RNU12*	Code le snRNA U12
		Syndrome de Lowry-Wood	LWS	Douleurs articulaires Malformation des pieds/mains/genoux Nanisme Microcéphalie	*RNU4atac*	Code le snRNA U4atac
		Syndrome de Roifman	RFMN	Nanisme Microcéphalie Retard intellectuel Dystrophie rétinienne Anomalies squelettiques Dimorphisme faciale Immunodéficience	*RNU4atac*	Code le snRNA U4atac
		Syndrome de Taybi-Linder	TALS	Nanisme sévère Microcéphalie sévère Retard intellectuel Dimorphisme faciale Dysplasie squelettique Eczéma	*RNU4atac*	Code le snRNA U4atac

A. Mutations de protéines associées au spliceosome

Les protéines associées aux spliceosomes sont les facteurs d'épissages externes, qui n'entrent pas dans la composition des spliceosomes et qui servent principalement à déterminer quels sites d'épissage seront utilisés, et internes, qui sont associés aux snRNA et forment les snRNP.

a. Facteurs d'épissages externes

Nous pouvons citer trois exemples de maladies liées à une mutation sur un facteur d'épissage : la dystrophie musculaire des ceintures de type 1G (LGMD1G), la cardiomyopathie dilatée (DCM) et la sclérose latérale amyotrophique (ALS) ou maladie de Charcot. Toutes ces maladies altèrent le fonctionnement des muscles, plus spécifiquement ceux de la ceinture pelvienne et scapulaire pour LGMDIG et le muscle cardiaque pour DCM, alors que l'ALS affecte l'ensemble des muscles squelettiques. Nous nous intéresserons plus particulièrement à cette dernière, de par ses liens avérés avec l'épissage mineur et son incidence relativement élevée (2 à 3 malades pour 100000 personnes par an) (Hardiman et al., 2017), l'ALS étant la maladie touchant les motoneurones la plus commune chez l'adulte (Jutzi et al., 2018).

Les gènes en cause pour la LGMD1G et la DCM sont *HNRPDL* et *RBM20*, tous deux des facteurs d'épissage. Les cibles de HNRPDL ne sont pas connues chez l'Homme, la seule étude sur cette protéine ayant été effectuée chez la levure et le poisson-zèbre (Vieira et al., 2014). RBM20 a pour cible divers gènes impliqués dans le développement cardiaque, dont la Titin, une protéine abondante dans les muscles striés et est un élément essentiel à leur fonctionnement.

L'ALS est une maladie causant la mort de motoneurones, neurones terminaux de la transmission d'un signal nerveux effectuant une action sur l'organe, le muscle ou la glande ciblée par une impulsion motrice. Ses principales caractéristiques sont une rigidité, un tremblement, un rétrécissement et une faiblesse grandissante des muscles, provoquant notamment des difficultés pour marcher, parler, avaler et respirer. Ces symptômes apparaissent le plus souvent entre 50 et 75 ans et provoquent la mort en 2 à 5 ans (Jutzi et al., 2018). Dans plus de 90 % des cas, la cause de l'ALS est inconnue et aucun traitement n'est proposé à ce jour. Dans ~10 % des cas, l'ALS est provoquée par des mutations souvent localisées dans les régions de faible complexité des gènes *FUS* et *TARDBP* (*TDP-43*) (Scotti and Swanson, 2016), ou encore dans les gènes *SOD1* et *C9orf72*, qui n'encodent pas de facteur d'épissage (Jutzi et al., 2018). Dans la majorité des cas, ces mutations provoquent l'ubiquitination et l'accumulation cytoplasmique de la protéine TARDBP (Jutzi et al., 2018). Les protéines FUS et TARDBP peuvent reconnaître des motifs riches en GU présents dans des milliers de gènes, dont d'autres facteurs d'épissage, permettant de réguler l'épissage dans le système nerveux central. Plus spécifiquement, la protéine TARDBP est impliquée dans la régulation de l'expression de transcrits avec de grands introns, dans l'épissage d'exons spécifiques, et dans la répression de l'épissage de sites cryptiques introniques (Scotti and

Swanson, 2016). La protéine TARDBP co-localise avec une structure nucléaire adjacente aux cajal bodies, les GEMini of cajal bodies (GEM), et serait impliquée dans leur formation, comme le propose des expériences de KO de TARBP dans des cellules de souris et des cellules humaines. Les GEM ont aussi la particularité d'être trouvés avec la plus grande densité et la plus grande taille dans les motoneurones. Des observations similaires ont été faites avec la protéine FUS : elle interagit avec les snRNA et peut réduire leur niveau nucléaire en les piégeant dans le noyau, elle interagit avec la protéine SMN qu'on retrouve aussi au niveau des GEM et elle semble permettre la formation des GEM (Jutzi et al., 2018). Plusieurs études rapportent une réduction du niveau de di-snRNP U11/U12 dans les lymphocytes, la moelle épinière et le cortex moteur, et des défauts de formation du tri-snRNP U4atac/U6atac.U5, ainsi que des défauts d'épissage mineur, mais encore une fois de manière tissu-spécifique (Vazquez-Arango and O'Reilly, 2018).

La suppression de TARDBP a un effet direct, mais tissu-spécifique, sur la quantité de snRNA mineurs, notamment dans la moelle épinière, le cortex moteur et le thalamus. De plus, il est parfois observé, dans les neuromoteurs de la moelle épinière, une réduction de la quantité de la protéine 59K, qui est spécifique au snRNP U11. Concernant la protéine FUS, elle interagit préférentiellement avec les ARNm contenant un intron et avec le spliceosome mineur, un KO de FUS dérégulant l'épissage et l'expression de certains gènes U12 (Jutzi et al., 2018). Malheureusement, il n'existe pas, à ce jour, d'études transcriptomiques focalisées sur l'épissage mineur dans l'ALS, qui pourraient permettre de déterminer la contribution du spliceosome mineur dans cette maladie neurodégénérative.

b. Protéines associées aux snRNA

A notre connaissance, il existe six pathologies liées à des mutations des protéines des snRNP, dont cinq pourraient affecter l'épissage majeur et mineur, la dernière provoquant un défaut du spliceosome mineur uniquement.

La dystrophie mendibulo-faciale-microcéphalie (MFDM) et le syndrome de Burn-McKeown (BMKS) sont provoqués par des mutations dans les gènes *MFTUD2* et *TXNL4A*, tous deux codant une protéine associée au seul snRNA U5 partagé entre le spliceosome majeur et mineur, nécessaire à l'action catalytique des spliceosomes. Le MFDM et le BMKS ont chacun des caractéristiques spécifiques, le premier provoquant une microcéphalie et un retard intellectuel et le second provoquant une surdité et des défauts cardiaques/rénaux, mais ils partagent une dystose mandibulo-faciale. On retrouve cette caractéristique couplée à une

malformation des membres dans le syndrome de Nager (NAS), provoqué par des mutations dans le gène *SF3B4*, qui code une protéine du snRNP U2 et U12 se fixant sur le BPS et donc impliquée dans la reconnaissance des introns à épisser. Ces trois maladies présentent des caractéristiques remarquablement semblables, alors que deux étapes distinctes de l'assemblage des spliceosomes sont affectées. Pourtant, d'autres mutations sur des gènes codant des protéines associées à U2 uniquement (*U2AF*) ou à U2 et U12 (*SF3B1*, *ZRSR2*) provoquent toutes trois une maladie extrêmement différente, le syndrome myélodysplasique (MDS), caractérisé par une leucémie myélomonocytaire chronique (Vazquez-Arango and O'Reilly, 2018). La protéine ZRSR2 a un rôle particulier et important dans le spliceosome mineur, où elle est impliquée dans la reconnaissance du site 3' d'épissage. En conséquence, le transcriptome de patients MDS avec mutation sur *ZRSR2* présente des rétentions d'introns U12 et d'introns U2 adjacents, ainsi que l'activation de sites d'épissage cryptique de type U2 (Madan et al., 2015). Enfin, d'autres mutations affectant le snRNP U5 ou U4/U6 et U4atac/U6atac provoquent une maladie encore une fois radicalement différente : la rétinite pigmentaire (RP). La RP cause la dégénérescence progressive de photorécepteurs dans la rétine et n'affecte aucun autre tissu (Vazquez-Arango and O'Reilly, 2018). Ces altérations de l'épissage mènent donc à des phénotypes étonnamment tissu-spécifique, compte tenu de l'action globale des spliceosomes.

Une seule maladie affectant une protéine d'un snRNP spliceosome mineur-spécifique a été identifiée à ce jour (Tableau I). Elle correspond à une forme de déficience en hormone de croissance isolée (IGHD) due à des mutations dans le gène *RNPC3* encodant la protéine 65K spécifique au di-snRNP U11/U12 (Argente et al., 2014). Un des transcrits mutés forme un PTC et est dégradé par le NMD, réduisant l'abondance de 65K (Norppa et al., 2018). Les patients IGHD présentent un retard de croissance post-natal sévère, un retard de maturation des os et une microcéphalie légère, vraisemblablement liés à l'incapacité des cellules à produire un di-snRNP U11/U12 fonctionnel. Pour cette maladie, une étude transcriptomique sur la qualité de l'épissage des introns mineurs a été menée, sur des cellules sanguines mononucléées (deux patients vs. quatre contrôles). Ces expériences ont mis en évidence une liste de 21 gènes U12 dans lesquels les introns mineurs étaient mal épissés chez les patients, parfois coordonnés avec une rétention des introns U2 flanquant l'intron U12, ainsi que des sauts d'exons et des événements d'épissage dans lesquels des sites cryptiques U2 adjacents aux sites d'épissage mineurs étaient préférentiellement utilisés chez les patients (par exemple,

dans le gène *SPCS2*, Figure 11). Une sur-expression inattendue et inexpliquée, deux fois supérieure aux contrôles, a aussi été observée pour le snRNA U4atac (Argente et al., 2014).

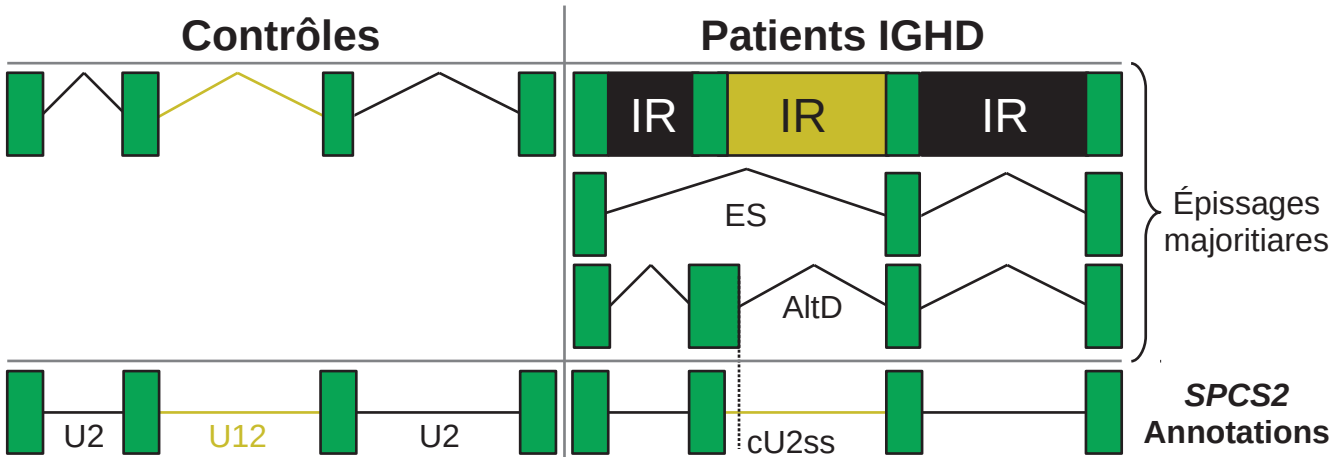

Figure 11 : Formes d'épissage majoritaires observées chez les patients IGHD et leurs contrôles pour le gène *SPCS2*.
Seuls les 4 derniers exons et les 3 derniers introns de *SCSP2* sont schématisés. Les exons, introns U2 et U12 sont colorés en vert, noir et jaune, respectivement. cU2ss : site d'épissage U2 cryptique, IR : Intron Retention, ES : Exon Skipping, AltD : Alternative Donor. *Inspiré de (Argente et al., 2014).*

Chez le zebrafish, un étude suggère que la protéine 65K pourrait aussi avoir un rôle dans des stades de l'assemblage du spliceosome postérieur à la reconnaissance de l'intron et pourrait interagir avec l'ARNsn U6atac (Verma et al., 2018).

B. Défauts des snRNA

Les maladies liées à des défauts dans les snRNA peuvent avoir deux origines : des mutations dans des gènes codants les protéines impliquées dans leur maturation ou des mutations dans les gènes des snRNA eux-mêmes.

a. Mutation des protéines de maturation

Nous pouvons citer cinq maladies liées à un défaut de maturation des snRNA, dont une liée au complexe Integrator (INT), une liée à un nouveau mécanisme de maturation et trois liées au complexe SMN. Parmi ces dernières, l'amyotrophie spinale (SMA), à l'image de l'ALS, a des liens établis avec l'épissage mineur et représente la maladie du motoneurone la plus fréquente chez l'enfant (Jutzi et al., 2018), avec une incidence d'environ 1 malade pour 10000 naissances par an (Sugarman et al., 2012).

Des mutations dans deux gènes du complexe Integrator (INT), *INTS1* et *INTS8*, provoquent un syndrome non-caractérisé que nous nommerons déficience en INT (INTD) et provoquant une

déficience mentale sévère, une incapacité de parler, des difficultés motrices, des épilepsies, une microcéphalie légère, une hypoplasie du pont et du bulbe rachidien, un défaut de migration des neurones corticaux et des malformations faciales et des membres. Dans les cellules des patients, l'INT est défaillant et mène à une accumulation de snRNA non-matures (prouvé pour U1, U2 et U4) et à des variations importantes du profil d'expression et d'épissage des gènes (Oegema et al., 2017).

L'hypoplasie pontocérébelleuse de type 7 (PCH7) est caractérisée par une dégénérescence des neuromoteurs, une atrophie ou hypoplasie du pont et du bulbe rachidien, des faiblesses musculaires et des anomalies de respiration. Le PCH7 est causé par des mutations dans le gène *TOE1* qui code une déadenylase ayant la particularité d'être concentrée dans les Cajal Bodies et impliquée dans la maturation des snRNA (prouvé uniquement pour U1, U2 et U5). Elle semble être l'exonucléase responsable de la dégradation de l'extrémité 3' des snRNA après la coupure effectuée par l'INT. Cette fonction est surprenante, sachant que TOE1 cible les queues polyA des ARNm *in vitro* et que les snRNA n'ont pas de queue polyA connue, bien que diverses observations portent à croire qu'ils sont polyadénylés, ce qui faciliterait le recrutement de TOE1 sur ces snRNA non-matures (Lardelli et al., 2017).

Le syndrome cérébrocostomandibulaire (CCMS) et l'hypotrichose simple (HS) sont causés par des mutations sur *SNRPB* et *SNRPE*, deux gènes codants des protéines Sm s'assemblant sur tous les snRNA exportés dans le cytoplasme. Bien que ces protéines fassent parties du même complexe, les phénotypes des syndromes sont radicalement différents : le premier provoque principalement des défauts du développement des côtes, de la mâchoire et de la langue (ressemblant à ceux du MFDM, BMKS et du NAS) ; le second conduit à une perte des cheveux, poils, cils et sourcils. Dans ces maladies, le niveau de snRNP parait globalement inchangé (Vazquez-Arango and O'Reilly, 2018).

L'amyotrophie spinale (SMA) est causée par des mutations dans le gène *SMN1*, codant la protéine SMN essentielle à la formation du Sm ring sur tous les snRNA (sauf U6 et U6atac) dont la quantité est réduite chez les patients. Comme l'ALS, le SMA conduit à la mort des motoneurones et à une atrophie des muscles et la protéine SMN est aussi enrichie dans les GEM (Jutzi et al., 2018). Bien que la maturation des snRNP majeurs et mineurs devrait être perturbée dans cette maladie, des études semblent indiquer que le spliceosome mineur est particulièrement affecté : le niveau de snRNP mineur est réduit dans les tissus touchés ; la formation du tri-snRNP U4atac/U6atac.U5 est altérée dans des lymphoblastes ; de nombreux

défauts d'épissage mineurs sont observés à la fois chez les patients SMA, mais aussi dans des modèles drosophile et souris (Jutzi et al., 2018).

Les ressemblances entre ALS et SMA, tant au niveau du phénotype que des dérèglements du spliceosome mineur, semblent indiquer que cette machinerie d'épissage joue un rôle spécifique et crucial dans les neurones (Jutzi et al., 2018; Vazquez-Arango and O'Reilly, 2018). Cette hypothèse est aussi soutenue par les maladies causées par des mutations sur les gènes des snRNA mineurs.

b. Mutation des snRNA

Chez l'Homme, aucune mutation des gènes codant les snRNA majeurs n'est associée à une pathologie, ce qui peut être dû à la présence de nombreuses copies de ces gènes dans le génome, ou bien indiquer que de telles mutations soient incompatibles avec la vie. Cependant, chez la souris, une mutation d'une copie du gène codant le snRNA U2 provoque une neurodégénérescence progressive affectant particulièrement le cervelet, laissant penser que l'expression des copies des snRNA pourrait être régulée de manière tissu-spécifique (Jia et al., 2012). Parmi les quatre maladies présentées dans le Tableau I, *i.e.* l'ataxie cérébelleuse (CA), le Syndrome de Lowry-Wood (LWS), le Syndrome de Roifman (RFMN) et le Syndrome de Tayb-Linder (TALS), nous nous intéresserons particulièrement à LWS, RFMN et TALS, toutes trois des maladies autosomiques récessives affectant le snRNA U4atac, sujet d'étude principal de cette thèse.

L'ataxie cérébelleuse (CA) est causée par une mutation homozygote sur le gène *RNU12* codant le snRNA U12 et est caractérisée par des faiblesses musculaires menant à des difficultés pour s'asseoir et marcher, liées à une hypoplasie et une dégénérescence du développement du cervelet. Les cellules sanguines mononucléées de ces patients présentent un niveau élevé de snRNA U12 et des rétentions d'intron U12 globales (trouvées par RT-qPCR et RNA-seq) parfois couplée avec une dérégulation de l'expression du gène U12 (Elsaid et al., 2017).

Trois syndromes distincts sont causés par des mutations homozygotes ou hétérozygotes composites (deux allèles présentant deux mutations différentes) dans le gène *RNU4ATAC* codant le snRNA U4atac : le syndrome de Lowry-Wood (LWS) (Farach et al., 2018), le syndrome de Roifman (RFMN) (Merico et al., 2015) et le syndrome de Taybi-Linder (TALS ou MOPDI) (Edery et al., 2011; He et al., 2011). Ces syndromes sont caractérisés par une microcéphalie, une déficience intellectuelle, un retard de croissance et des malformations

osseuses, mais ils ne partagent pas les mêmes particularités faciales. De plus, ces syndromes ont chacun des spécificités : le LWS présente des problèmes oculaires (Farach et al., 2018), le RFMN, en plus d'anomalies de la rétine, est caractérisé par une immunodéficience (Merico et al., 2015) et le phénotype des TALS est le plus sévère, menant à une mort précoce (souvent avant trois ans) et inexpliquée des patients (Edery et al., 2011). A ce jour, 4, 10 et 30 familles, comportant 5, 14 et 53 cas de LWS, RFMN et TALS respectivement, ont été identifiées dans le monde, représentant 23 mutations sur *RNU4ATAC* (Figure 12) (Bogaert et al., 2017; Dinur Schejter et al., 2017; Farach et al., 2018; Ferrell et al., 2016; Hallermayr et al., 2018; Heremans et al., 2018; Lionel et al., 2018; Putoux et al., 2016; Shaheen et al., 2019; Shelihan et al., 2018; Wang et al., 2018).

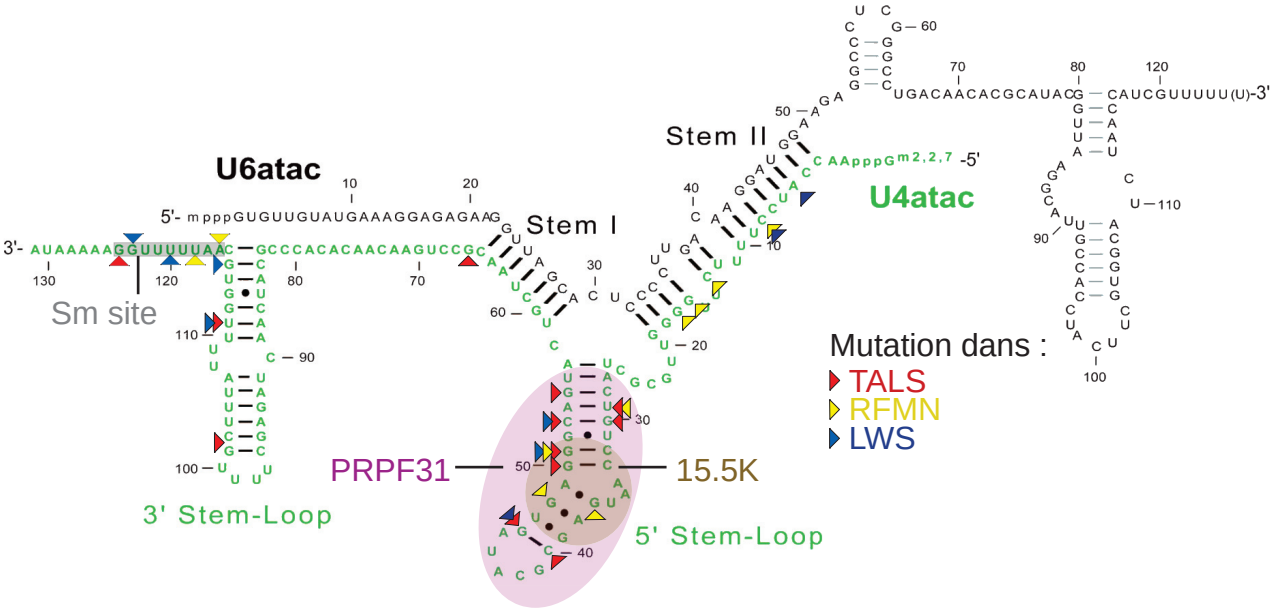

Figure 12 : Mutations identifiées dans les pathologies affectant le snRNA U4atac.
U4atac (vert) est représenté en association avec le snRNA U6atac (noir). Les mutations trouvées dans le TALS, RFMN et LWS sont représentées par des triangles rouges, jaunes et bleus, respectivement. Un triangle entre deux acides nucléiques indique une insertion. Les sites de fixation aux protéines PRPF31, 15.5K et Sm sont indiqués en violet, marron et gris, respectivement. *Adapté de (Edery et al., 2011).*

Quelques particularités concernant ces mutations peuvent être notées. Les mutations provoquant le TALS ne sont jamais situées dans la tige II (stem II) et les cas les plus sévères de cette maladie sont associés à la forme homozygote du variant le plus fréquemment identifié, le g.51G>A. A l'opposé, le RFMN implique toujours une mutation dans la stem II associée de manière hétérozygote composite à un variant TALS-like, c'est-à-dire extérieur à la stem II (une exception existe pour un patient RFMN g.16G>A homozygote). Le LWS est à la frontière des deux précédents syndromes, puisque les combinaisons de mutations qu'il

présente peuvent ressembler à ceux des TALS ou des RFMN. Les mutations dans la tige-boucle 5' (5' stem-loop) déstabilisent vraisemblablement les interactions ARN-protéines cruciales pour le fonctionnement de U5 (Schneider et al., 2002). Une corrélation entre la sévérité du TALS et l'affinité du di-snRNA U4atac/U6atac avec la protéine 15.5K semble exister, puisque les mutations perturbant le plus cette interaction causent les phénotypes les plus sévères de ce syndrome (Abdel-Salam et al., 2012; Jafarifar et al., 2014). Les mutations situées à l'extrémité 3' de U4atac se localisent sur le site de fixation des protéines Sm, ce qui pourrait corrompre la maturation du snRNA. L'ensemble de ces mutations provoquerait une malformation du tri-snRNP U4atac/U6atac.U5, à l'exception du variant g.124G>A qui réduirait l'abondance du snRNA U4atac (Jafarifar et al., 2014).

Le RFMN est le syndrome le mieux caractérisé des trois d'un point de vue transcriptomique, puisque trois analyses RNA-seq de cellules sanguines mononucléées, de mégacaryocytes ou de cellules sanguines périphériques d'un total de six patients RFMN ont été publiées à ce jour. Toutes caractérisent une rétention globale des introns U12 spécifiquement, et deux rapportent également une sur-expression des snRNA (Figure 13). Il est cependant étonnant que des snRNA soient détectés alors que, pour le séquençage des ARN, les ARNm ont été préalablement sélectionnés en capturant leur queue polyA, queue que n'ont théoriquement pas les snRNA. Cette observation rejoint celle faite pour PCH7 et semble indiquer que les snRNA sont polyadenylés à un moment de leur cycle de maturation. Pour le LWS, il n'existe aucune étude transcriptomique. Concernant le TALS, l'équipe de Patrick EDERY, une des deux équipes à avoir identifiée *RNU4ATAC* comme gène causal de la maladie et équipe dans laquelle j'ai effectué ma thèse, a également mis en évidence des rétentions d'intron U12 par RT-qPCR dans des fibroblastes de patient (Edery et al., 2011). Dans le but de produire une étude robuste et globale du transcriptome dans le contexte de ce syndrome, et ainsi caractériser de manière précise les défauts d'expression et d'épissage conséquents à des mutations sur *RNU4ATAC*, l'équipe a collecté, pendant 15 années et à travers le monde, des cellules de patients TALS pour en produire la plus grande cohorte établie à ce jour. Par l'utilisation du séquençage de l'ARN (RNA-seq) et l'utilisation de méthodes d'analyse bioinformatique, que nous décrirons dans le chapitre suivant, nous avons pu dresser le profil transcriptomique des patients TALS grâce aux travaux réalisés durant cette thèse.

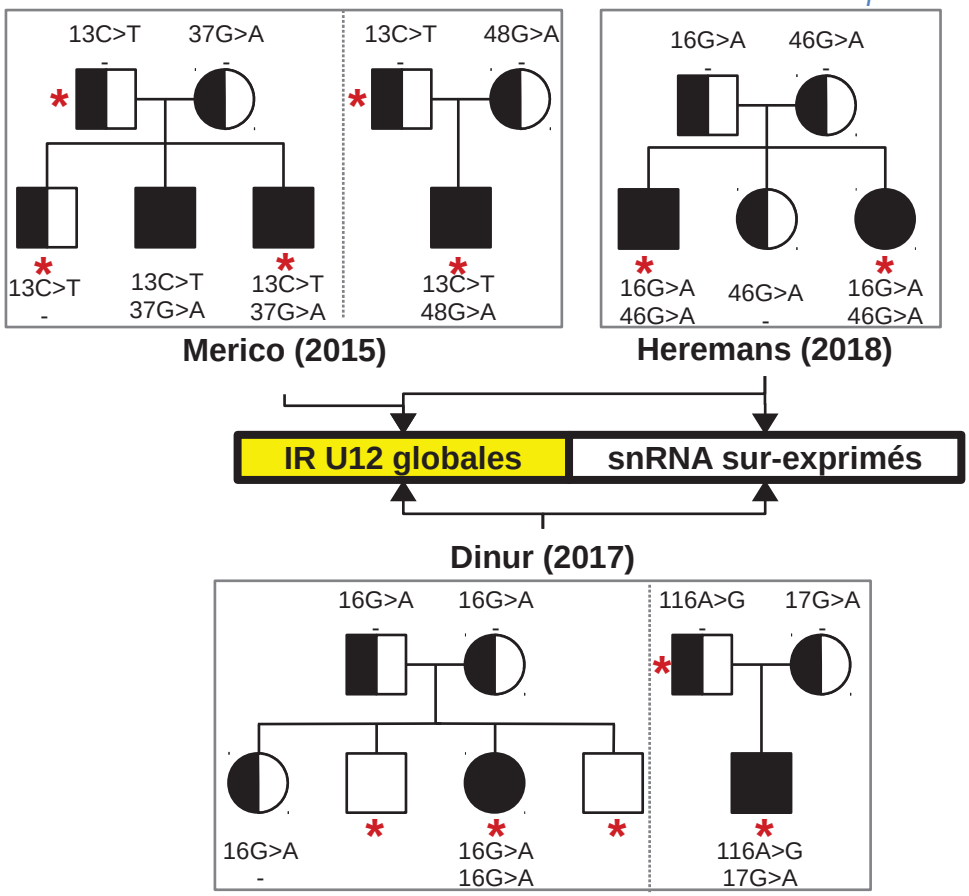

Figure 13 : Familles, patients RFMN et contrôles analysés en transcriptomique associés aux principaux résultats de l'analyse.
Les individus féminins et masculins sont représentés par des ronds et des carrés, respectivement. Les individus homozygotes ou hétérozygotes sains et homozygotes ou hétérozygotes composites affectés sont indiquées avec du blanc et en noir, respectivement. Si elles existent, les mutations sur *RNU4ATAC* sont indiquées sous les individus. Une astérisque indique les individus séquencés. *Adapté de (Dinur Schejter et al., 2017; Heremans et al., 2018; Merico et al., 2015).*

III. Analyse du transcriptome par RNA-seq

Dans ce chapitre, nous discuterons de la méthode utilisée durant cette thèse pour accéder à l'information transcriptomique des cellules (RNA-seq), ainsi que des outils et stratégies bioinformatiques employés pour détecter des anomalies de l'expression des gènes ou de l'épissage des ARN-pre. L'expression des gènes sera ici définie comme la quantité d'ARN correspondant à un gène (production moins dégradation) mesurée à un instant donné.

A. RNA-seq (Illumina)

Le RNA-seq est une technologie consistant à déterminer la composition nucléotidique de séquences d'ARN (Mortazavi et al., 2008). Depuis 2005, le RNA-seq a apporté une contribution capitale à la recherche et à la médecine, puisque cette technique rend l'analyse du transcriptome rapide et peu coûteuse, sans demander de connaissance *a priori* sur le transcriptome séquencé, et est aujourd'hui la norme pour les analyses à large échelle. Bien qu'il existe plusieurs méthodes pour réaliser un RNA-seq, nous ne présenterons ici que celle utilisée lors de cette thèse : le séquençage Illumina.

Le RNA-seq peut être divisé en 3 étapes : préparation des banques d'ADN complémentaires (ADNc), l'amplification de l'ADN et le séquençage de l'ADN.

a. Préparation des librairies

Cette première étape doit établir quels types d'ARN devront être séquencés, et donc quels ARN sélectionner. Le séquençage direct de l'ARN étant une technique récente qui n'égale pas encore la maîtrise du séquençage de l'ADN, les ARN doivent être rétro-transcrits en ADNc avant d'être séquencés. Pour conserver précisément la séquence de l'ARN dans l'ADNc, il faut au préalable fragmenter les ARN.

Suivant l'étude, il peut être intéressant de se restreindre à certains types de tissus et à certaines populations d'ARN, e.g. les petits ARN ou les ARNm. Comme nous l'avons vu dans les chapitres précédents, les profils transcriptomiques peuvent être radicalement différents selon les tissus, rendant cette étape cruciale pour l'analyse des données générées et l'interprétation des résultats. L'ARN et l'ADN de cellules du tissu sélectionné sont alors extraits puis traités avec une désoxyribonucléase, qui va réduire la quantité d'ADN présent dans la solution. La quantité et qualité de l'ARN sont ensuite évaluées, puis soumises à un enrichissement de la population d'ARN souhaitée. Pour cette thèse, les ARN étudiés sont les ARNm. Une des

techniques d'enrichissement des ARNm peut se faire par l'ajout de billes magnétiques couvertes de séquences polyT auxquelles vont s'hybrider la queue polyA des ARNm. Tous les transcrits sans queue polyA (notamment les ARN non-codants ou les ARNm de gènes d'histones) ne sont théoriquement pas séquencés.

Les ARN sont ensuite fragmentés, généralement en chauffant la solution d'ARN, puis sélectionnés en fonction de leur taille, pour optimiser la qualité de la rétro-transcription. En effet, de longs ARN ont tendance à former des structures secondaires, empêchant la revers-transcription, et l'enzyme utilisée pour cette étape a aussi tendance à « s'essouffler » et à faire de plus en plus d'erreurs au cours de la transcription de l'ARN en ADNc. Des adaptateurs, qui seront utilisés lors de l'étape d'amplification et qui peuvent servir à identifier les fragments appartenant à un même échantillon, sont fixés aux extrémités des ARN fragmentés. Des amorces composées de six nucléotides aléatoires, permettant théoriquement d'échantillonner toutes les positions des ARN, et une reverse-transcriptase sont alors ajoutées pour convertir les ARN en ADNc.

b. Amplification

L'amplification a pour but de synthétiser suffisamment d'ADN pour produire, à l'aide de fluorochromes, un signal lumineux détectable par les machines Illumina, lors du séquençage de l'ADNc. Pour Illumina, cette amplification se fait par réaction en chaîne grâce à une polymérase (polymerase chain reaction, PCR), schématisée dans la figure 14.

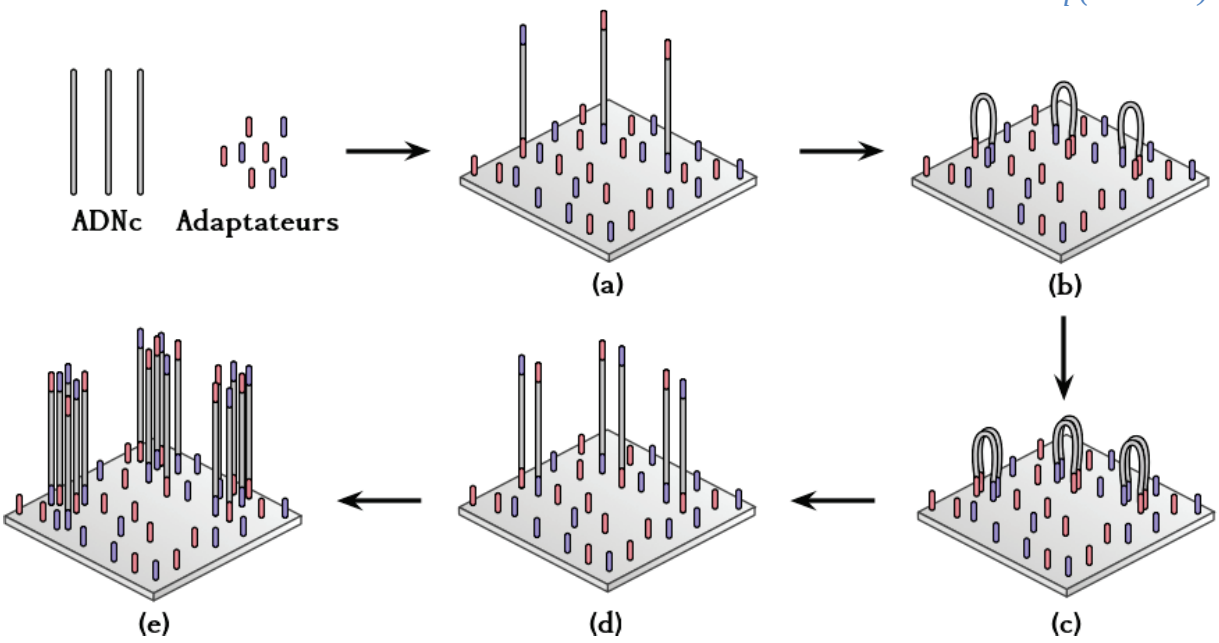

Figure 14 : Amplification par ponts.
Les étapes (a), (b), (c), (d) et (e) sont décrites dans le texte. *Source : https://www.atdbio.com/content/58/Next-generation-sequencing.*

A l'aide d'un des adaptateurs, les ADNc vont se fixer, par complémentarité de séquence, à une plaque sur laquelle sont ancrés des milliers d'oligonucléotides (a). L'étape de génération de clusters commence alors et va aboutir à plusieurs milliers de copies de chaque fragment d'ADNc. Pour cela, l'adaptateur libre des ADNc va lui aussi se lier à une séquence de la plaque, formant ainsi un pont d'ADNc (b), ce qui permettra d'initier la synthèse du brin complémentaire avec l'ajout d'une ADN polymérase et de nucléotides (c). Les brins sont alors dénaturés (d) et ces étapes sont répétées plusieurs fois, permettant l'amplification locale de chaque molécule d'ADNc (e).

c. Séquençage

Le séquençage Illumina se fait par synthèse du brin complémentaire des séquences des clusters, à l'aide de l'ADN polymérase, d'amorces et de nucléotides modifiés. Ces nucléotides sont des terminateurs réversibles de la synthèse (un seul nucléotide peut être ajouté à un brin complémentaire par cycle) et sont marqués par fluorescence, une couleur correspondant à chaque base. Les amorces, constituées de séquences aléatoires, seront les points de départ des ADN polymérases qui ajouteront les nucléotides modifiés. A chaque inclusion, une image du signal fluorescent de la plaque est prise, permettant de déterminer quelle base a été ajoutée dans chaque cluster. Le terminateur et le fluorochrome du nucléotide ajouté sont alors retirés, et un nouveau cycle commence pour déterminer les bases suivantes (Figure 15). Le séquençage s'arrête lorsqu'un certain nombre de bases séquencées est atteint (de 75 à 300 en

fonction des machines). Les séquences déterminées pour chaque cluster formeront une lecture (read). Dans la majorité des utilisations actuelles, ce séquençage se fait de manière paired-end, c'est à dire que les deux extrémités des ADNc sont séquencées.

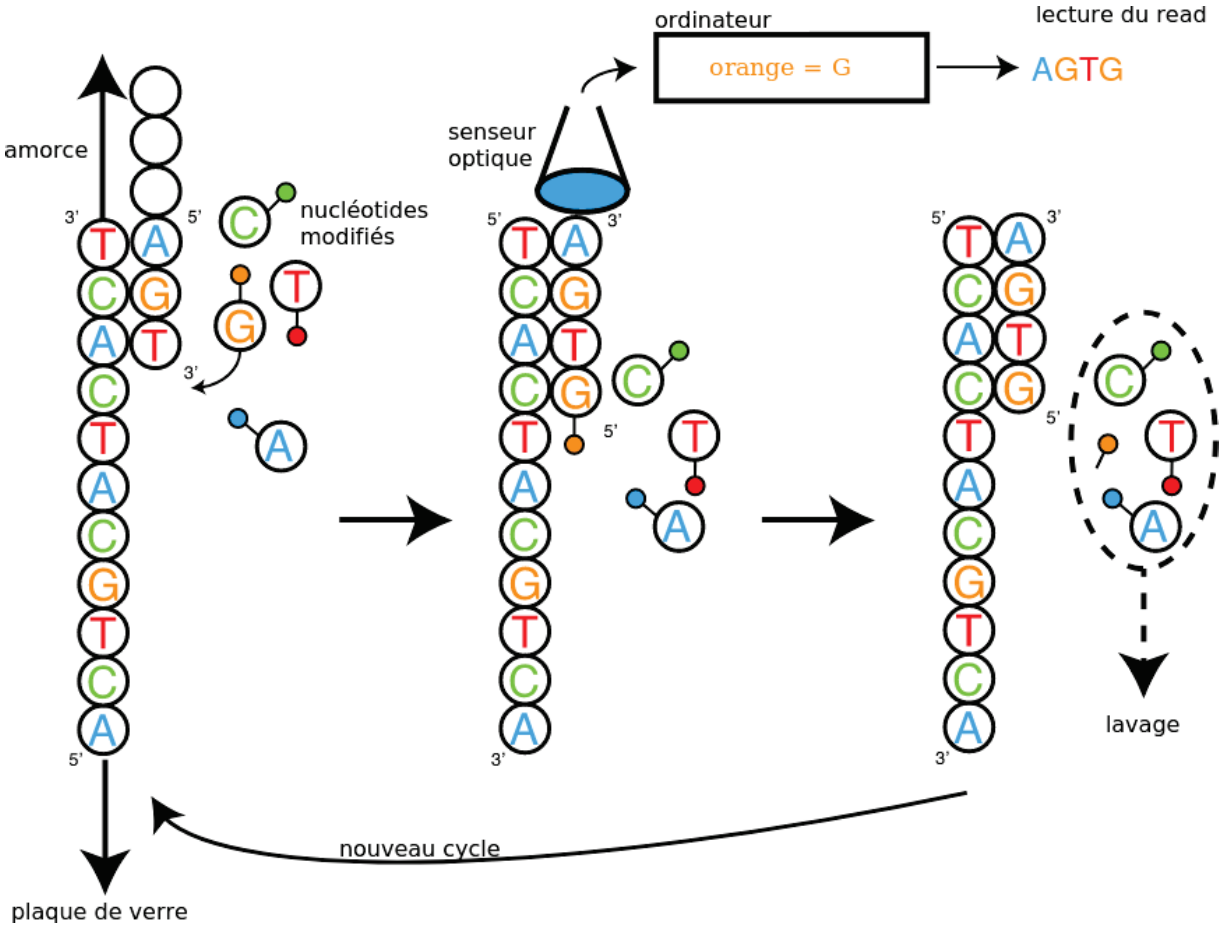

Figure 15 : Séquençage Illumina.
Les principales étapes nécessaires à la lecture de la séquence d'un read sont représentées. *Source : https://binf.snipcademy.com/lessons/ngs-techniques/illumina-solexa*.

Le séquençage Illumina permet de produire plusieurs milliards des reads de très bonne qualité pour un coût relativement faible, faisant de cette technologie la plus utilisée sur le marché. Il existe cependant des biais de séquençage, pouvant intervenir à chaque étape, les plus connus étant liés aux amorces aléatoires lors de la reverse-transcription (Hansen et al., 2010) et au contenu en GC des reads (Ross et al., 2013). De plus, les étapes de PCR pourraient favoriser l'amplification de certains ADNc, tout en altérant les quantités relatives des ARN sélectionnés. Les fichiers de sortie d'un séquençage sont écrits dans un format particulier, le FastQ, qui, en plus de donner la séquence de chaque read, associe une valeur de qualité à chaque nucléotide, permettant de juger de leur qualité et, par extension, de celle du read. Les fichiers FastQ peuvent être plus au moins volumineux suivant la question biologique adressée et les conditions expérimentales (~5GB/FastQ dans notre étude).

d. Contrôle qualité

Lors de l'obtention de reads, la première étape à réaliser est le contrôle qualité. Plusieurs outils sont disponibles pour réaliser ce contrôle, le plus utilisé étant probablement FastQC. Cette étape va permettre de mettre en évidence diverses anomalies dans un jeu de données, comme des reads de mauvaise qualité qu'il faudra filtrer, la présence de séquences adaptatrices qu'il faudra couper ou encore des séquences sur-représentées qu'il faudra analyser pour s'assurer que le jeu de données n'a pas été contaminé. Cette étape est nécessaire à n'importe quelle analyse de données RNA-seq, pour nettoyer le jeu de données et améliorer les performances des analyses qui en découleront.

e. Planification expérimentale

La valeur de profondeur (nombre de reads unique contenant un nucléotide donné) choisie pour le séquençage aura un impact direct sur la quantité de reads produits mais aussi sur le coût du séquençage. Il convient d'adapter cette profondeur en fonction de la question biologique d'une expérience, afin de ne pas sur- ou sous-séquencer, le premier constituant une perte d'argent et le second empêchant de produire une analyse approfondie. Par exemple, pour une analyse de l'épissage, les transcrits alternatifs peuvent ne représenter qu'une petite fraction de l'expression d'un gène (Liu et al., 2013), séquencer trop peu conduira ainsi à la perte de ces transcrits alternatifs rares.

Il est aussi important de prendre en compte l'espace disque occupé par les données générées, un fichier correspondant à un échantillon pouvant prendre une place de plusieurs GB, qu'il faudra multiplier par deux pour un séquençage paired-end. La technologie permet de séquencer toujours plus profondément plus d'échantillons dont le stockage des données brut est un problème informatique majeur en bioinformatique. De plus, pour les analyses différentielles, il est indispensable de posséder des réplicats techniques (re-séquencer le même échantillon biologique) ou mieux, des réplicats biologiques (séquencer plusieurs échantillons biologiques distincts pour une même condition expérimentale) pour produire une analyse des échantillons qui reflétera au mieux la population. Il est important d'estimer quelle place sera prise par l'ensemble de ces données et de bien réfléchir au nombre de replicats et à la profondeur de séquençage nécessaire pour aboutir à une utilisation raisonnée des ressources informatiques et biologiques.

Les milliard de reads générés devront ensuite être analysés. Devant cette énorme quantité de données, il est indispensable de posséder une certaine expertise en informatique et biologie, la

première pour produire une analyse en profondeur et en un temps raisonnable et la seconde pour décider de l'analyse la mieux adaptée au contexte biologique et interpréter correctement les résultats.

B. Analyse en composante principale

Avant de réaliser une analyse différentielle, il peut être utile d'explorer visuellement les résultats de séquençage, pour contrôler leur qualité, se familiariser avec les jeux de données et/ou pour donner des indications sur les types d'analyses à effectuer.

Avec des données de RNA-seq, nous pouvons par exemple nous intéresser au niveau d'expression des gènes (modélisé par les Transcript Per Milion, voir III.C.a.Expression), qui peut être variable d'une condition expérimentale à l'autre et qui permettrait ainsi d'identifier des groupes d'échantillons au profil d'expression distinct. Pour détecter les gènes à l'expression anormalement variables entre deux groupes, une analyse différentielle pourra être réalisée, dans un second temps. Dans un premier temps, il faut pouvoir identifier les échantillons avec un profil d'expression variable : plus la variabilité sera grande, meilleure sera la discrimination entre échantillons.

Une matrice de comptage est fréquemment utilisée pour sauvegarder les niveaux d'expression de chaque gène (en ligne) dans chaque échantillon (en colonne). Chaque gène représente alors une dimension des données. Avec plus de 10 000 gènes exprimés, il n'est pas possible de représenter l'intégralité de l'information contenue dans une matrice de comptage. Une des technique couramment utilisée pour permettre une représentation approximative des données est l'analyse factorielle, méthode de la famille des statistiques multivariées, dont une des analyses proposées est l'Analyse en Composante Principale (ACP).

L'ACP vise à identifier et hiérarchiser les dimensions (les gènes dans notre exemple) ou combinaisons de dimensions, appelées composantes (PC), qui représentent le mieux la variabilité des données. Il est ensuite possible de ne s'intéresser qu'aux premières composantes (expliquant les plus grands pourcentages de variance entre échantillons) et de les visualiser par paires (PC1 vs PC2 / PC1 vs PC3 etc.), souvent à l'aide d'un nuage de points. Par exemple, la Figure 16 représente la composante 1 et 2 de l'ACP du niveau d'expression des gènes dans 13 échantillons, provenant de 3 types cellulaires différents : des fibroblastes, des amniocytes et un lymphoblaste. Ces échantillons ont été colorés, *a posteriori*, en fonction de leur type cellulaire. La PC1, représentée sur l'axe des abscisses, explique 30 % de la

variabilité totale des données et sépare principalement l'échantillon lymphoblaste de tous les autres. La PC2, représentée sur l'axe des ordonnés, explique 23 % de la variabilité totale des données et sépare principalement l'échantillon lymphoblaste des échantillons amniocytes. Une combinaison des PC1 et PC2 (segment noir) sépare aussi nettement les échantillons fibroblastes des échantillons amnyocites. L'ACP regroupe ainsi les échantillons par leur type cellulaire, indiquant que le profil d'expression est plus variable entre ces groupes qu'entre les échantillons d'un même tissu. En tout, ce graphique représente seulement 53 % de la variabilité totale des données : il peut donc être intéressant de visualiser d'autres composantes pour voir si de nouveaux groupes se forment (par exemple, en fonction du sexe).

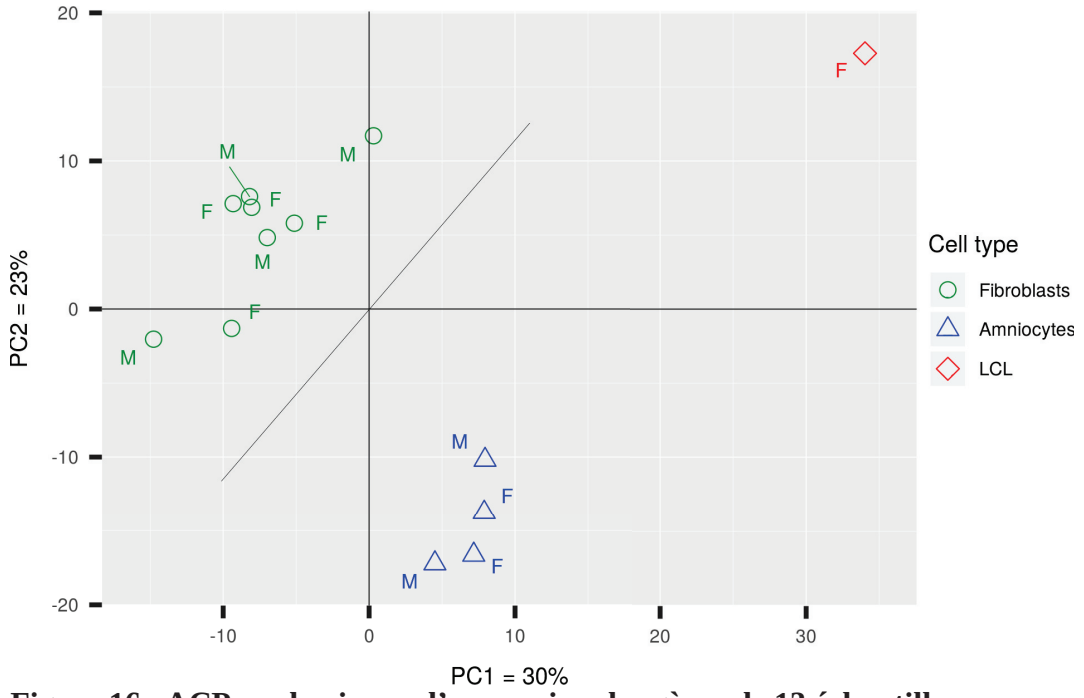

Figure 16 : ACP sur le niveau d'expression des gènes de 13 échantillons.
Le sexe de l'individu sur lequel a été prélevé l'échantillon est indiqué à proximité des points (M : Homme, F : Femme). Les échantillons verts, bleus et rouge proviennent de fibroblastes, amniocytes et lymphoblastes, respectivement. Un segment noir symbolise la combinaison des axes x et y séparant les fibroblastes des amniocytes.

L'ACP peut aussi être utilisée pour discriminer les échantillons sur d'autres bases que leur niveau d'expression respectif. Par exemple, dans le cadre de cette thèse, nous nous intéresserons principalement à la qualité de l'épissage des introns (modélisé par le Percent Spliced In (PSI), voir III.C.b.Épissage), et plus particulièrement des introns U12. Nous devons donc modifier la matrice de comptage, pour qu'elle représente les PSI de chaque intron (en ligne) dans chaque échantillon (en colonne).

L'ACP permet également une identification de points aberrants (outliers) n'appartenant à aucun groupe, échantillons qui pourront éventuellement être exclus des analyses postérieures pour en améliorer la qualité.

C. Analyses différentielles

Les analyses différentielles comparent des données provenant de deux groupes pour détecter et quantifier des variations entre ces groupes, nommées effets. Chaque membre d'un même groupe présente au moins une même caractéristique qui n'est pas partagée avec les membres de l'autre groupe, nommé condition expérimentale (*i.e.* homme/femme, traité/non-traité, patient/contrôle).

Le but de l'analyse différentielle est de tester une hypothèse nulle, le plus souvent du type « Il n'existe pas de variation entre les deux groupes », en calculant la probabilité d'obtenir les valeurs observées si cette hypothèse est vraie : la p-value (Tenny and Abdelgawad, 2019). Pour rejeter l'hypothèse nulle, il faut que cette probabilité tombe en-dessous d'un certain seuil, le plus généralement 0.05 (on parle alors de test significatif). Ce seuil correspond au risque α : c'est la probabilité de rejeter une hypothèse nulle alors qu'elle est vraie (et donc, la probabilité de détecter des faux-positifs). Parallèlement, le risque β correspond à la probabilité d'accepter une hypothèse nulle alors qu'elle est fausse (et donc, la probabilité de détecter des faux-négatifs). La puissance d'un test (sa capacité à détecter des vrais-positifs) est donc 1- β. Pour une même comparaison, réduire l'un de ces seuils conduit irrémédiablement à l'augmentation du second : réduire le nombre de faux-positifs augmentera le nombre de faux-négatifs et inversement. Pour réduire le risque α et β, il faudra augmenter la taille de son échantillon, ou recueillir les valeurs observées à l'aide d'une méthode plus précise pour réduire la dispersion entre les valeurs observées.

Si le test statistique est répété sur plusieurs observations, un pourcentage de 5 % de faux-positifs (risque α) sera attendu lors de chaque test, ce qui peut mener au final à un nombre élevé de faux-positifs : plus il y a de comparaison, plus il est probable d'obtenir un résultat positif par chance. Si les tests individuels sont indépendants, des méthodes existent pour parer ce problème. L'une d'elle est la correction de Bonferroni, qui va modifier le seuil associé au risque α en le divisant par le nombre de tests effectués. Si beaucoup de tests sont réalisés, cette correction est très conservative : on aura très peu de faux-positifs au prix de nombreux faux-négatifs, ce qui réduit la puissance statistique. Cette méthode est à privilégier dans le cas où un seul résultat faux-positif est problématique. Un méthode alternative, disposant d'une plus

grande puissance statistique et consistant à contrôler le taux de faux-positifs, est la procédure de Benjamini-Hochberg. Un seuil Q doit être choisi pour indiquer le taux de faux-positif acceptable. Les m p-values sont classées par ordre décroissant et un rang i leur est attribuée (la p-value la plus haute est à i=1, la deuxième plus haute est à i=2 etc... jusqu'à la plus faible qui est à i=m). La p-value la plus élevée qui satisfait l'équation p-value < (i/m)Q est considérée significative, comme toutes les autres p-value qui lui sont de rang supérieur. Des p-value ajustées (ou FDR) sont parfois calculées avec la formule suivante : p-value ajustée = p-value * m/i. Si la p-value ajustée est inférieure à Q, le test est considéré comme significatif (McDonald, 2015).

L'utilisation de la p-value est parfois critiquée par la communauté scientifique, le plus souvent à cause du biais de publication, qui décrit la perception biaisée de la recherche : les expériences avec un résultat positif (significatif) ont beaucoup plus tendance à être publié que celles avec un résultat négatif (non-significatif) (Mlinarić et al., 2017). Grossièrement, si 100 études sont menées sur des événements indépendants avec un seuil de significativité de 5 %, 95 expériences auront un résultat négatif, probablement non publié (effet tiroir), et 5 auront un résultat positif lié au hasard, qui aura plus de chance d'être publié. Ce biais est particulièrement problématique lors de méta-analyses. Pour le contrer, les journaux scientifiques peuvent encourager la réplication des résultats ainsi que la publication des résultats négatifs ; il est aussi possible de contraster un résultat positif avec l'ampleur de son effet (voir paragraphe suivant).

Le nombre de replicats biologiques est un paramètre important : son augmentation améliore la précision de la modélisation des variations biologiques dans chacun des groupes, permettant de discerner plus facilement des effets significatifs réellement liés à une condition (vrai-positifs) de ceux erronément attribués à la condition (faux-positifs). En augmentant le nombre de replicats biologiques, on améliore la modélisation de la variabilité entre les échantillons d'un même groupe et il est donc moins probable d'attribuer un effet lié à une caractéristique imprévue différenciant les groupes comparés : un facteur de confusion. Par exemple, un gène pourra paraître différentiellement exprimé en comparant deux patients masculins à deux contrôles féminins, mais finalement se révéler invariablement exprimé en ajoutant un patient féminin et un contrôle masculin à la comparaison. Le différentiel d'expression semblait être lié à la pathologie, alors qu'il était lié à un facteur de confusion : le sexe. Comme les facteurs de confusion sont très souvent trop difficiles à prévoir, seule l'augmentation du nombre d'échantillon dans les groupes étudiés permet de réduire ou abolir leur impact sur une analyse.

Avec un très grand nombre de replicats, des effets significatifs extrêmement faibles peuvent être détectés (Lin et al., 2013). Il est donc essentiel de quantifier l'effet (calculer son ampleur) pour se faire une idée de son intérêt. Globalement, moins il y aura de replicats, plus il y aura de vrai- et/ou faux-positifs associés à des effets forts, et plus il y aura de replicats, moins il y aura de faux-positifs associés à des effets forts, et plus il y aura de vrai-positifs associés à des effets faibles. Il est donc important d'établir des seuils sur l'ampleur des effets pour tenter de filtrer un maximum de faux-positifs et de conserver uniquement les vrai-positifs pertinents. Le résultat d'une analyse différentielle doit toujours être présenté avec l'ampleur de la différence.

En RNA-seq, la majorité des analyses vise à déterminer quels sont les gènes différentiellement exprimés entre deux conditions, et nous présenterons le pipeline conventionnellement suivi dans ce but. Il est aussi possible de conduire une analyse différentielle de l'épissage des transcrits entre deux conditions (qui peut aussi être vue comme un différentiel d'expression des transcrits alternatifs). Nous ne tenterons pas ici de comparer les méthodes d'analyses de l'expression ou de l'épissage (reviews : (Schurch et al., 2016; Soneson and Delorenzi, 2013; The RGASP Consortium et al., 2013a, 2013b)) ; nous nous contenterons de présenter celles utilisées lors de la thèse.

a. Expression

Le pipeline de détection et quantification des gènes différentiellement exprimés (gènes DE) peut être divisé en trois étapes : 1) Alignement des reads ; 2) Quantification du nombre de reads par gène ; 3) Modélisation de l'effet et test statistique.

1) Alignement

Cette première étape a pour but de déterminer de quelle partie du génome chaque read est originaire. Pour cela, il faut disposer d'un génome de référence et d'un logiciel d'alignement. Les quantifications ultérieures vont grandement dépendre de la qualité de cet alignement, qui doit aussi être suffisamment rapide pour aligner plusieurs millions de reads en un temps raisonnable. En RNA-seq, le travail de l'aligneur est d'autant plus dur qu'un read peut provenir d'une jonction exon-exon, avec un bout de sa séquence correspondant à un exon, et l'autre bout correspondant à un autre exon, séparés par un intron dans le génome de référence. Deux stratégies sont utilisées pour résoudre rapidement le problème de l'alignement : la méthode exon-first, utilisée par TopHat (Trapnell et al., 2009) et TopHat2 (Kim et al., 2013), et la méthode seed-and-extend, utilisée par STAR (Dobin et al., 2013) et HISAT2 (Kim et al., 2015) (Figure 17). La première stratégie consiste à tenter d'aligner entièrement tous les reads

sur le génome. Les reads alignés permettront d'identifier les exons et de construire tous les sites d'épissage possibles entre exons voisins, ceux-ci devant être séparés par une séquence génomique commençant et finissant par des di-nucléotides donneurs et accepteurs caractéristiques d'un intron. Les reads non-alignés seront alors réalignés, en utilisant cette fois la liste des jonctions exon-exon construite à l'étape précédente comme référence. La deuxième stratégie, seed-and-extend, commence par couper les reads en petits morceaux appelés graines. L'algorithme construit également une carte d'un ensemble de fragments du génome. Les graines sont alignées efficacement sur ces fragments génomiques, ce qui permet de déterminer rapidement les provenances possibles d'un read. L'alignement de chacune de ces graines est ensuite étendu jusqu'à couvrir le read complet. Pour ces deux méthodes, il est aussi possible d'utiliser des annotations comme guides.

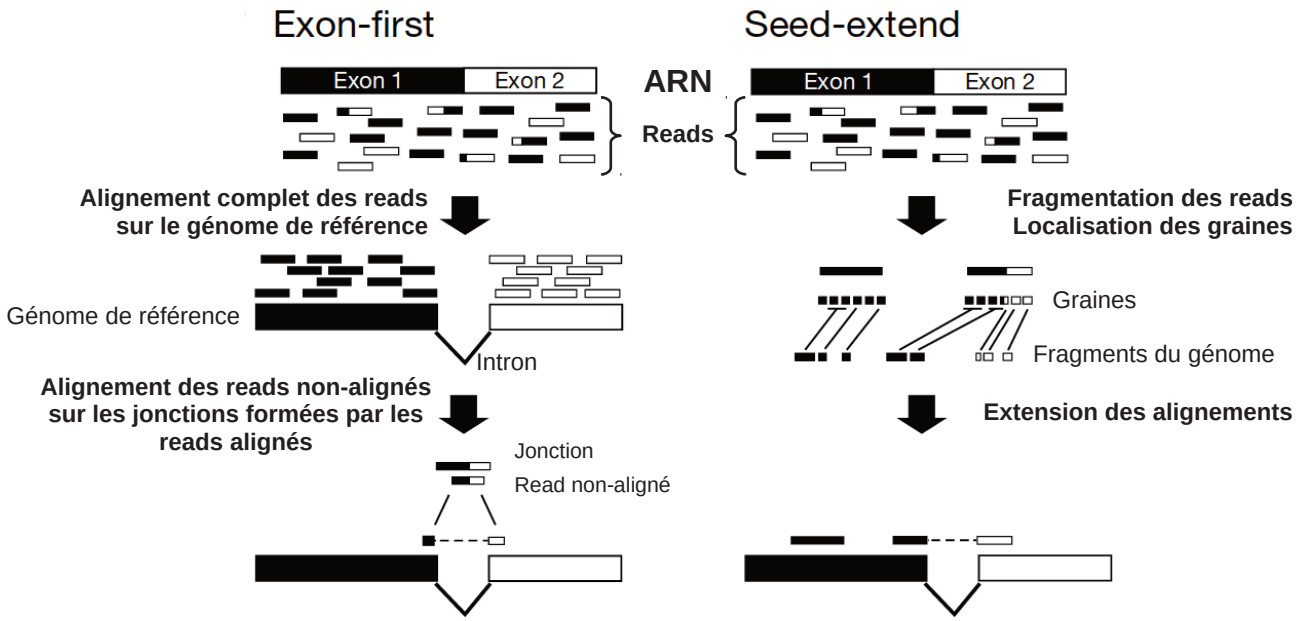

Figure 17 : Algorithmes pour aligner des reads provenant de RNA-seq.
Les méthodes exon-first et seed-and-extend sont présentées. *Adapté de (Garber et al., 2011).*

En plus de la difficulté de résoudre rapidement le problème de l'alignement des reads, de nombreux cas particuliers obligent les aligneurs à faire certains choix. Par exemple, si un read s'aligne parfaitement à deux endroits dans le génome, l'aligneur pourra préférer ne rapporter aucun alignement, ou rapporter aléatoirement l'un d'eux comme optimal, et l'autre comme secondaire. Puisque la plupart des logiciels d'analyse de fichiers d'alignement n'utilisent que les alignements uniques, leur puissance statistique sera réduite pour certaines régions répétées du génome (Robert and Watson, 2015).

2) Quantification de l'expression des gènes

Dans un premier temps, il faut déterminer le nombre de reads prouvant l'expression de chaque gène, une étape pouvant être réalisée avec HTSeq-count (Anders et al., 2015). Cette étape produira un tableau avec tous les gènes de l'annotation en ligne et le comptage de reads provenant de ces gènes en colonne, pour chaque échantillon.

Il est ensuite possible de passer à l'étape suivante, ou de normaliser les comptages obtenus pour établir une métrique décrivant l'expression d'un gène de manière comparable avec les autres gènes de la même condition ou d'une condition différente. Les paramètres qui ont un impact important sur le nombre de reads généré pour un gène sont la profondeur de séquençage de l'échantillon et la taille des gènes. La métrique la plus utilisée est le RPKM ou FPKM (Reads ou Fragments Per Kilobase per Milion mapped reads), le premier si le séquençage est single-end, le deuxième si le séquençage est paired-end. Pour calculer le R/FPKM, il faut : a) Calculer le facteur de mise à l'échelle (PM : Per Million scaling factor) : somme des reads d'un échantillon / 1 000 000 ; b) Normaliser pour la profondeur de séquençage (RPM, Reads Per Million) : comptage / PM (pour chaque gène) ; c) Normaliser pour la taille des gènes (RPKM) : RPM / taille du gène en kilobase (pour chaque gène). Cependant, il est difficile de comparer les R/FPKM provenant de différentes conditions, puisque leur somme pour chaque condition sera différente. Ainsi, une métrique plus adaptée et équivalente est souvent préférée : le TPM (Transcripts Per Million). Pour calculer les TPM, il faut utiliser les même étapes précédentes mais dans un ordre différent : a) Normaliser les comptages par la taille des gènes (RPK, Reads Per Kilobase) : comptage / taille du gène en kilobase (pour chaque gène) ; b) Calculer le PM : somme des RPK / 1 000 000 ; c) Normaliser par la profondeur de séquençage (TPM) : RPK / PM (pour chaque gène). Les TPM présentent ainsi l'avantage d'avoir une somme égale à un million dans chaque échantillon, facilitant grandement leur comparaison. A l'aide de ces valeurs, il est possible de réaliser des analyses descriptives comme l'ACP.

3) Modélisation et analyse statistique

Nous décrirons ici seulement la méthode utilisée par DESeq2 (Love et al., 2014), l'un des outils les plus utilisés pour la détection de gènes différentiellement exprimés (DE). Cet outil va tester, pour chaque gène, l'hypothèse nulle qu'il n'y a pas de différence entre les niveaux d'expression des deux conditions, c'est à dire que l'expression du gène n'est pas affectée par la condition. Pour cela, DESeq2 va modéliser les comptages, estimer les Fold-Change

Logarithmiques (LFC, magnitude de l'effet) dans le cadre du modèle linéaire généralisé et tester l'hypothèse nulle avec le test de Wald.

Pour des données de RNA-seq, le nombre de reads attribué à chaque gène est décrit à l'aide d'une distribution de Poisson ou d'une distribution Binomiale Négative (NB), suivant l'utilisation de replicats techniques ou biologiques, respectivement. Par rapport à la distribution de Poisson, la NB ajoute un paramètre de sur-dispersion permettant de mieux modéliser les variations biologiques entre les replicats (Di et al., 2011). La moyenne de cette distribution sera proportionnelle au nombre de fragments provenant d'un gène d'un échantillon, normalisée par la profondeur de séquençage de l'échantillon. La variabilité entre replicats sera estimée à l'aide d'un paramètre de dispersion, crucial pour l'analyse différentielle mais souvent imprécis pour des analyses avec peu de replicats biologiques. DESeq2 prend en compte ce problème et fait l'hypothèse que des gènes à l'expression moyenne similaire ont des valeurs de dispersion similaires. Cette heuristique permet à DESeq2 d'être performant, même dans le cas ou peu de replicats biologiques sont disponibles (Schurch et al., 2016).

Les LFC, logarithme du ratio de l'expression d'un gène entre deux conditions, permettent de quantifier la magnitude de l'effet : plus la valeur absolue du LFC est élevée, plus l'expression du gène varie fortement. Cependant, le LFC étant un ratio, ces valeurs sont extrêmement variables pour des gènes avec de faibles comptages. DESeq2 va au préalable filtrer les gènes associés à de faibles comptages (qui ont de toute façon peu de chance d'être détectés comme différentiellement exprimés), et réduire l'estimation des LFC vers zéro d'autant plus fortement que l'expression ou la dispersion d'un gène est faible ou élevée, respectivement. Ce filtre et cette réduction des LFC offre une quantification plus reproductible des différences d'expression entre les gènes.

b. Épissage

La détection et l'analyse différentielle de l'épissage sont des tâches globalement plus difficiles que l'analyse différentielle de l'expression. Pour réaliser une analyse exhaustive, il est recommandé de séquencer profondément le transcriptome pour détecter les transcrits alternatifs rares, d'utiliser une taille de read d'au moins 100 nucléotides pour détecter plus facilement les jonctions exon-exon.

Il existe deux niveaux de distinction pour les outils dédiés à l'analyse de l'épissage : les méthodes basées sur l'alignement (mapping-first) ou l'assemblage (assembly-first), toutes deux pouvant être des méthodes locales ou globales.

Les méthodes mapping-first, couramment utilisées pour les organismes modèles, nécessitent un génome et un transcriptome de référence, le résultat de l'analyse dépendant grandement de leurs qualité. Comme pour l'analyse des gènes DE, la première étape est l'alignement des reads. Les méthodes assembly-first sont principalement utilisées pour des organismes ne possédant pas un génome de référence de bonne qualité, bien qu'elles trouvent aussi leur utilité chez les organismes modèles. Les étapes de ces méthodes sont l'assemblage des reads en contigs, ce qui est indépendant de toute référence, puis l'alignement des contigs sur un génome de référence, si celui-ci est disponible. Comme montré pour les sauts d'exon, les approches mapping-first et assembly-first sont complémentaires, la première permettant de détecter plus de transcrits alternatifs faiblement couverts et la seconde permettant de détecter plus de transcrits alternatifs utilisant des jonctions d'épissage non-annotées, ce qui est fréquent même pour les organismes modèles (Benoit-Pilven et al., 2018). Cette dernière propriété peut s'avérer très utile pour l'analyse de l'épissage dans le contexte de maladies liées au spliceosome mineur, puisque nous avons vu que des sites cryptiques ou *de novo* étaient parfois utilisés.

Les méthodes locales cherchent à identifier les événements d'épissage et à quantifier les formes d'épissage qui les composent. Au contraire, les méthodes globales tentent de reconstruire et quantifier les transcrits alternatifs complets (variants d'épissages), problème particulièrement complexe avec des reads courts et très partiellement solvable (The RGASP Consortium et al., 2013b). Nous nous focaliserons ici sur les méthodes locales.

Les méthodes locales peuvent quantifier les événements d'épissage de différentes façons. Nous prendrons ici l'exemple d'une rétention d'intron (IR) (Figure 18, haut). Pour déterminer le nombre de reads spécifique à chaque forme d'épissage, ces méthodes vont utiliser la jonction entre les exons constitutifs de l'événement (exons A et B) ainsi que les jonctions entre un exon constitutif et un autre exon du gène (exons X) pour quantifier la forme d'exclusion (reads verts). Pour la forme d'inclusion, la partie variable (intron S) ainsi que les deux jonctions entre un exon constitutif et S (AS et SB) seront utilisés pour la quantification (reads jaunes/oranges). Pour considérer qu'un read atteste de la présence d'une jonction, il

faut qu'un certain nombre de ces bases s'alignent sur les deux composants de la jonction (on parle d'overhang, cinq bases alignées sont généralement demandées).

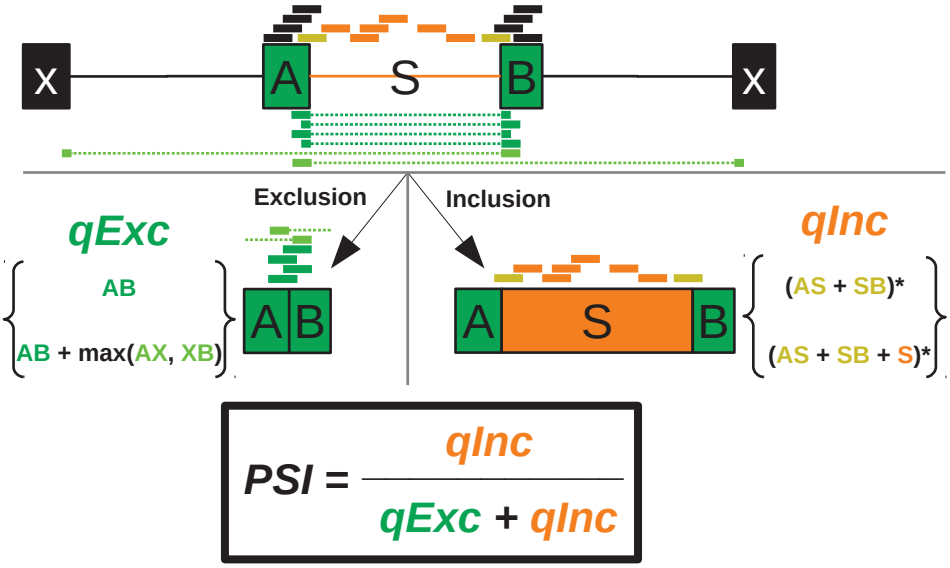

Figure 18 : Quantification d'un événement d'épissage : exemple d'une rétention d'intron.
Les exons constitutifs et la partie variable de l'événement d'épissage apparaissent en vert (exons A et B) et orange (intron S), respectivement. Les rectangles noirs (X) symbolisent tout autre exon du gène considéré. Les formes d'épissage sont représentées (partie basse de la figure) avec pour chacune l'ensemble des reads pouvant être utilisé pour leur quantification. Les reads sont colorés en vert foncé, vert clair et jaune si ils appartiennent aux jonctions AB, AX ou XB et AS ou SB, respectivement. Les reads introniques et exoniques sont colorés en orange et noir, respectivement. Un astérisque indique une normalisation par la taille de la partie exclue quantifiée par rapport à la partie incluse quantifiée. PSI : Percent Spliced In ; qExc : quantification de la forme d'exclusion ; qInc : quantification de la forme d'inclusion.

Différentes méthodes de calcul des quantifications peuvent être employées pour chacune des deux formes d'épissage (Figure 18, bas). Pour l'exclusion, on peut vouloir calculer le nombre de fois où les exons A et B sont reconnus (représenté par le nombre de reads sur AB), ou le nombre de fois où l'exon A ou B est reconnu (représenté par la somme du nombre de reads sur AB et AX ou XB). Pour l'inclusion, le mieux est d'utiliser les reads s'alignant à la fois sur la jonction AS et SB, mais il faut pour cela que la taille de la partie variable soit inférieure à la taille des reads, ce qui est très rare dans le cas de rétentions d'intron. Deux autres quantifications de la forme d'inclusion peuvent alors être employées. La première consiste à utiliser uniquement le nombre de reads sur les jonctions AS et SB, ce qui quantifie seulement le nombre de fois où les jonctions de S ne sont pas reconnues par le spliceosome, mais ne traduit pas forcément une rétention complète de la partie variable S. L'avantage de cette

méthode est qu'elle est indépendante des biais de quantification pouvant être trouvés dans les régions introniques (discuté dans la suite de ce chapitre), mais la puissance statistique sera amoindrie, puisqu'un nombre réduit de reads sera pris en compte. La deuxième méthode de quantification va ajouter les reads s'alignant dans S à AS et SB, quantifiant effectivement la rétention de l'intron. Comme la fenêtre génomique prise en compte pour quantifier la forme d'inclusion est plus grande que celle de la forme d'exclusion, la quantification de l'inclusion doit être normalisée pour la taille.

Une métrique permet de refléter l'abondance relative des formes d'épissage : le Percent Spliced In (PSI, Figure 18), calculé en faisant le ratio de la quantification de la forme d'inclusion (normalisé par la taille) sur la somme des quantifications des formes d'inclusion et d'exclusion. Le PSI traduit ainsi l'abondance que représente la forme d'inclusion par rapport aux deux formes d'épissage ; ainsi, des PSI de 0 %, 50 % et 100 % indiquent que la forme d'inclusion est absente, que les deux formes d'épissages sont équitablement abondantes et que la forme d'exclusion est absente, respectivement.

L'IR est un événement particulièrement complexe à quantifier (Figure 19) (Vanichkina et al., 2018). En effet, les introns humains sont souvent très longs, remplis de régions de faible complexité, notamment des séquences répétées de type Alu, et peuvent contenir diverses annotations, comme des gènes codant des petits ARN. Puisque les aligneurs préféreront ne pas rapporter d'alignements multiples, un déficit de couverture en reads sera observé au niveau des régions de faible complexité ; inversement, si des gènes introniques sont exprimés, un pic de couverture en reads sera observé à leur niveau dans l'intron (Figure 19, croix rouges). Ainsi, toutes les régions des introns ne reflètent pas la couverture de l'intron uniquement. De plus, un effet bien connu de l'alignement de reads provenant de RNA-seq est l'hétérogénéité de couverture : la couverture a tendance à ne pas rester constante et à faire des « vagues » lorsque de longues distances génomiques sont observées, ce qui est lié à des biais de séquençage. Pour la plupart des événements d'épissage, l'hétérogénéité de couverture est négligeable, mais pour la rétention d'intron cet effet est plus important et doit être pris en compte (Figure 19, points d'exclamations rouges et bleus). Des méthodes adaptées sont donc nécessaires pour quantifier précisément les IR.

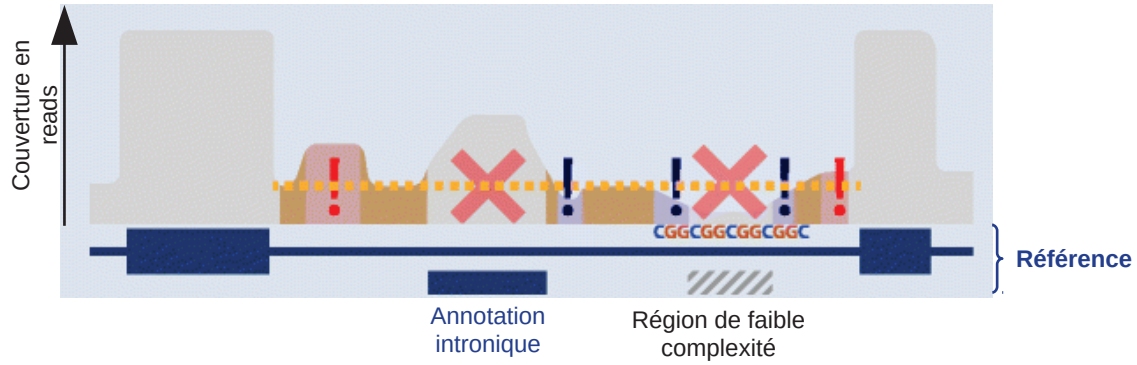

Figure 19 : Problèmes majeurs liés à la quantification des rétentions d'intron.
La couverture en reads (axe Y) de chaque nucléotide de la fenêtre génomique (axe X) est représentée. La moyenne de la couverture intronique est représentée par une ligne pointillée jaune. Dans la référence, les exons et les introns sont indiqués par des rectangles et des lignes bleues, respectivement. *Adapté de (Middleton et al., 2017).*

Nous prendrons ici l'exemple de trois méthodes d'analyse locale de différentiels d'épissage : 1) KisSplice (Sacomoto et al., 2012), assembly-first ; 2) vast-tools (Braunschweig et al., 2014; Irimia et al., 2014; Tapial et al., 2017), mapping-first, avec base de données d'événements d'épissage ; 3) IRFinder (Middleton et al., 2017), mapping-first, dédiée aux rétentions d'introns.

1) KisSplice, méthode assembly-first

KisSplice est un pipeline d'analyse différentielle de l'épissage comportant trois grandes briques : a) KisSplice (KS), qui assemble les reads, détecte et quantifie les événements d'épissage ; b) KisSplice2RefGenome (K2RG), brique optionnelle utilisant une référence pour annoter les événements d'épissage ; c) kissDE (Clara Benoit-Pilven, 2018), qui réalise l'analyse différentielle.

a) KS va couper l'ensemble des reads en k-mers, c'est-à-dire en mots de longueur k (41 par défaut), et les assembler pour former un Graphe de De Bruijn (GDB), ou chaque k-mer est un nœud, et chaque branche correspond à un chevauchement de longueur k-1 entre deux k-mers (Figure 20). Une structure particulière du GDB, appelée « bulle », sera toujours associée à deux variants, qui peuvent aussi bien correspondre à un événement d'épissage qu'à du polymorphisme ou des insertions/délétions. Une bulle est composée de deux nœuds constitutifs (au début et à la fin) et de deux chemins possibles pour relier ces nœuds, chaque chemin représentant ainsi un variant (Figure 20, chemin orange et vert). KS détecte l'ensemble des bulles et quantifie chaque chemin en y alignant les reads. L'exclusion sera toujours calculée en comptant les reads sur la jonction AB, et l'utilisateur peut choisir

d'utiliser la première ou deuxième méthode pour quantifier l'exclusion. Les paires de chemins et leur quantification respective sont écrites en sortie de KS.

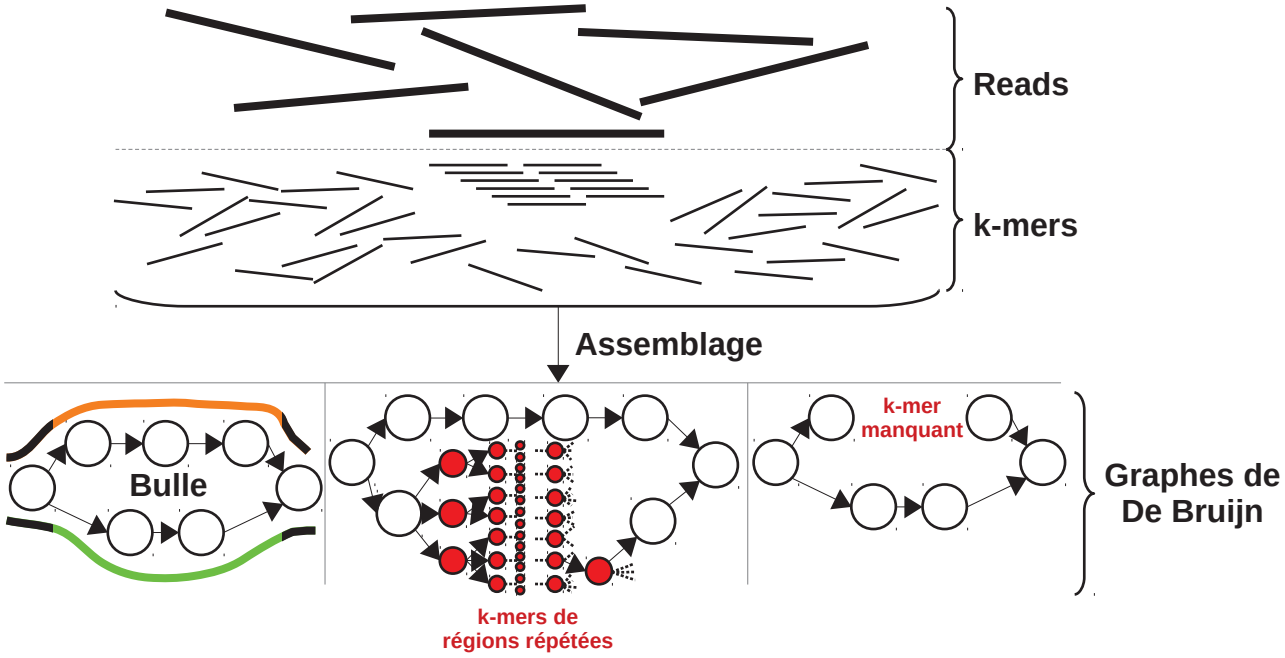

Figure 20 : Construction et assemblage des k-mers.
Chaque read est divisé en mot de longueur k (k-mer). Ces mots sont assemblés en prenant les k-mers comme nœuds (cercle) et un chevauchement de longueur k-1 comme arêtes (flèches reliant les cercles) pour former un Graphe de De Bruijn. Lorsqu'une arrête relie un nœud qui n'est pas représenté, des pointillés sont utilisés. Les bulles, structures particulières du graphe formant un chemin par variant (représenté en orange et vert), sont recherchées par KisSplice. Des régions particulièrement complexes ou des k-mers manquants dans le graphe empêchent la détection de certaines bulles.

Cette approche est peu efficace pour détecter des IR, et ce pour deux raisons (Figure 20). La première est liée aux régions de faible complexité : les reads provenant de ces régions vont générer des multitudes de k-mers interconnectés qui formeront des régions du GDB trop complexes pour être explorées efficacement par l'algorithme d'énumération des bulles, empêchant ainsi leur détection. La deuxième raison est liée à la grande taille des introns, à l'hétérogénéité de couverture et à la faible couverture en reads généralement observée dans les introns. Si une seule base de l'intron n'est pas couverte par un read, l'IR ne sera pas détectée puisqu'une bulle ne pourra pas être complètement formée dans le GDB : il manquera au moins un k-mer. Des solutions à ces problèmes peuvent être proposées, comme changer certains paramètres de KisSplice ou séquencer avec une plus grande profondeur, mais les approches mapping-first gardent tout de même un avantage certain pour la détection des IR.

L'intérêt principal de KS est son indépendance à toutes références, qui le place dans une position favorable pour détecter des formes d'épissage non-annotées (Benoit-Pilven et al., 2018). Comme nous l'avons vu dans le chapitre précédent, une déficience en spliceosome mineur peut mener à l'utilisation de sites cryptiques U2 autour des sites U12, ne menant plus à des IR mais à des sauts d'exon ou à des donneurs/accepteurs alternatifs. L'approche assembly-first de KS pourrait ainsi permettre de dénombrer de manière plus exhaustive ces événements particuliers.

b) Si l'utilisateur dispose d'un génome de référence, il peut aligner les chemins retournés par KS et utiliser le résultat de cet alignement avec K2RG pour annoter les événements d'épissage (Figure 21). Les chemins peuvent être beaucoup plus longs que les reads, ce qui permettra à un aligneur comme STARlong de détecter facilement les jonctions d'épissage. K2RG va principalement permettre de déterminer à quel type d'événement d'épissage alternatif appartient chaque couple de chemins et quelles sont les coordonnées génomiques de chaque site d'épissage. Si une annotation est utilisée avec K2RG, celui-ci déterminera aussi à quel gène appartient chaque événement.

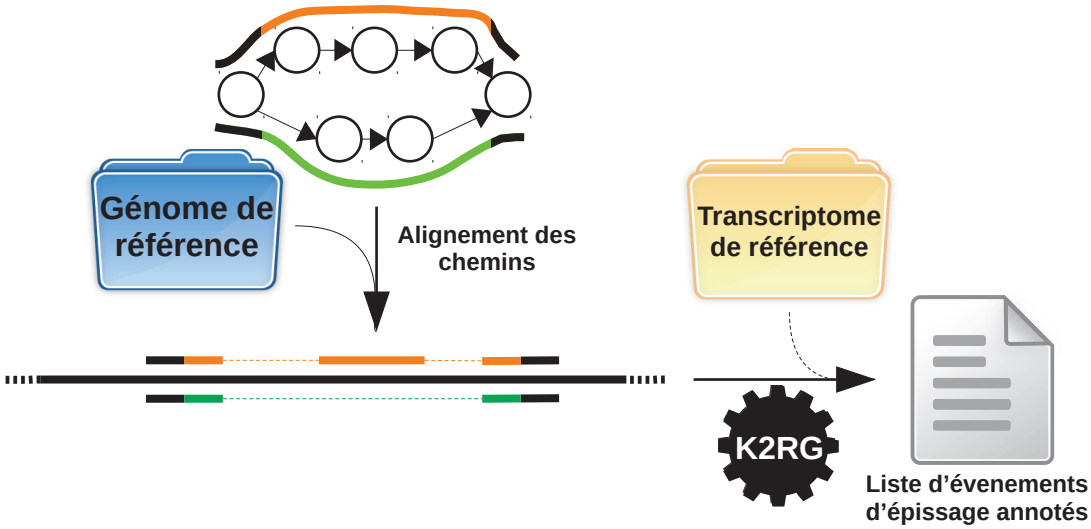

Figure 21 : Annotation d'événements d'épissage trouvés par KisSplice.
Pour chaque bulle, la séquence de chaque chemin est alignée sur un génome de référence (par exemple avec STARlong). L'alignement est ensuite traité par KisSplice2RefGenome (K2RG) pour créer un fichier texte tabulé représentant un événement d'épissage annoté par ligne. Un transcriptome de référence peut optionnellement être donné à K2RG pour ajouter des informations supplémentaires dans l'annotation.

c) Si l'utilisateur dispose de deux conditions biologiques et de replicats techniques ou biologiques, kissDE permettra de réaliser une analyse différentielle des événements d'épissage identifiés par KS. KissDE peut également être utilisé avec un tableau quantifiant

chaque forme d'épissage dans chaque échantillon produit par n'importe quel outil de détection et quantification des événements d'épissage, aussi bien assembly-first que mapping-first. Ce package bioconductor utilise des méthodes de normalisation et de modélisation semblables à DESeq2, mais adaptées au fait qu'il y a deux valeurs à prendre en compte pour l'épissage (quantification du variant d'inclusion et d'exclusion) contre une pour l'expression (quantification de l'expression du gène). Le test statistique réalisé par kDE permet de déterminer la probabilité que la variation de la qualité de l'épissage entre deux groupes soit liée à la condition expérimentale, en utilisant un test de maximum de vraisemblance.

2) Vast-tools, méthode mapping-first

La stratégie de vast-tools consiste à utiliser comme référence une liste finie d'événements d'épissage connue et d'en extraire la liste des jonctions exon-exon et d'introns (uniquement les 200 nucléotides centraux). Les régions de faible complexité de ces jonctions et introns sont détectées et ne seront pas prises en compte lors de la quantification. Les reads sont alors alignés en deux temps à l'aide de BOWTIE (Langmead et al., 2009), un aligneur qui ne tente pas d'aligner les reads sur des jonctions d'épissage. Le premier alignement se fait sur le génome, et le second, utilisant les reads non-alignés, se fait sur la liste prédéfinie des jonctions d'introns, permettant potentiellement de détecter toutes les IR, si l'intron est annoté. Cet outil ne permet pas l'identification de nouveaux événements d'épissage mais est particulièrement utile pour détecter des sauts d'exons de petites tailles, et octroie une attention particulière aux rétentions d'introns. En effet, vast-tools compare la couverture de l'intron à celle des jonctions exon-intron et intron-exon afin de s'assurer que l'intron ne présente pas de déséquilibre de couverture. Si tel est le cas, l'événement est exclu de l'analyse. La quantification de l'exclusion se fait en utilisant les reads sur la jonction AB, celle de l'inclusion se fait en utilisant les reads sur les jonctions AS et SB uniquement. Vast-tools propose également une analyse différentielle. Les développeurs de vast-tools ont aussi développé une base de données nommée VastDB, contenant l'ensemble des événements d'épissage détectés et quantifiés par leur outil sur une multitude d'échantillons humains, de souris et de poulet.

3) IRFinder, méthode mapping-first dédiée aux IR

IRFinder est un outil dédié aux rétentions d'intron, qui prend en compte l'ensemble des biais liés à leur quantification. IRFinder commence par extraire tous les introns d'un fichier d'annotation puis détermine les zones de ces introns qui ne seront pas prises en compte lors de la quantification. Comme pour vast-tools, les régions de faible complexité sont détectées en utilisant des reads synthétiques de 70 paires de bases, créés à partir du génome de référence, contenant une mutation en leur centre et séparés de 10 paires de bases les uns des autres. Ces reads seront alors alignés sur le génome de référence d'où ils proviennent. Toutes les 70 bases, IRFinder va observer combien de reads synthétiques se sont alignés de manière unique, et si moins de 5 reads sont trouvés, la région est exclue de la surface quantifiable de l'intron (Figure 22). Les zones introniques correspondant à des annotations ainsi que les cinq premiers et derniers nucléotides de l'intron (correspondant à l'overhang) sont également exclus. La taille efficace de l'intron est alors calculée en ne prenant en compte que les nucléotides conservés. Si cette taille est inférieure à 40 ou inférieure à 70 % de la taille d'origine de l'intron, celui-ci est exclu de l'analyse.

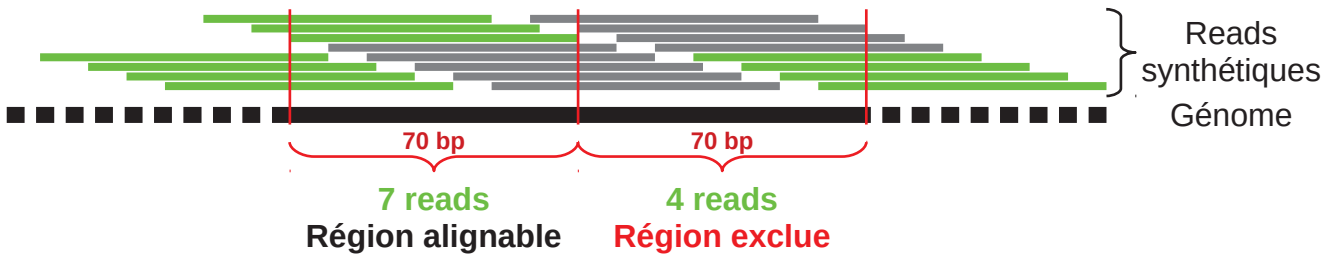

Figure 22 : Méthode d'identification des régions de faible complexité.
Les reads synthétiques alignés à un seul ou plusieurs endroits dans le génome sont indiqués en vert et gris respectivement. Les fenêtres génomiques de 70 nucléotides comportant moins de 5 reads uniquement alignés seront considérées comme des régions de faible complexité.

IRFinder utilise des fichiers d'alignements pour calculer la profondeur des introns, qui servira à quantifier la forme d'inclusion de l'IR (Figure 18, AS+SB+S). Pour prendre en compte l'hétérogénéité de couverture ou de potentiels gènes non-annotés, les 30 % des bases avec la plus haute et la plus basse couverture en reads sont aussi exclus du calcul de la profondeur de l'intron. La forme d'exclusion est quantifiée en utilisant les reads de jonctions AB et le maximum entre AX et XB. Cependant, le fichier de résultat de IRFinder contient aussi les quantifications de chaque jonction exon-exon et exon-intron.

Enfin, IRFinder propose plusieurs analyses différentielles. Si peu de replicats sont disponibles (entre 0 et 3), le test d'Audic et Claverie, traditionnellement utilisé pour détecter de rares différentiels d'expression, est préconisé. Sinon, le test de Student peut être utilisé, tout comme la méthode d'analyse différentielle de DESeq2 si des replicats biologiques sont disponibles.

c. Représentation d'événements d'épissage alternatif

Une tâche particulièrement compliquée lors de l'étude de l'épissage est la manière de visualiser les événements d'épissage alternatif et leur quantification.

La représentation la plus répandue est actuellement le Sashimi Plot (Katz et al., 2015). Ce type de visualisation permet de représenter un ou plusieurs évènements d'épissage situés dans une fenêtre génomique (Figure 23). Tout le long de cette fenêtre (axe X), la couverture en read de chaque base est représentée par un histogramme (axe Y). Les reads alignés sur une jonction exon-exon, et donc présentant un alignement discontinu sur le génome, sont représentés à l'aide d'un arc liant les deux bases jointes dans les reads. Au centre de ce trait est indiqué le nombre de reads alignés sur cette jonction. Le plus souvent, une représentation des transcrits connus est donnée tout en bas de la figure, ce qui permet d'identifier les exons et les introns, mais aussi de déterminer si des jonctions ou sites d'épissages sont nouveaux ou annotés.

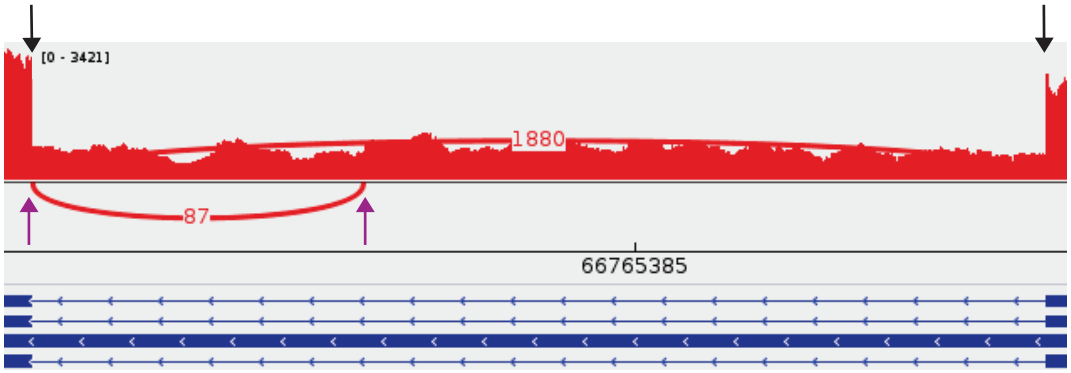

Figure 23 : Sashimi plot.
Représentation de la couverture en read (rouge) pour chaque base, l'échelle est donnée en haut à gauche. Les annotations des références apparaissent en bleu, les exons et introns sont représentés par des rectangles et des traits, respectivement, pour chaque transcrit annoté. Les reads alignés sur la jonction de deux bases éloignées sur le génome sont représentés par des traits horizontaux rouges, le nombre indique combien de reads sont observés sur cette jonction. Trois épissages sont présentés : l'épissage d'un intron long annoté (flèches noires), l'épissage d'un intron court non-annoté (flèches violettes) et la rétention de l'intron annoté (traduit par une forte couverture tout le long de l'intron).

Un Sashimi plot se construit à partir de données d'alignement, et est notamment réalisable par l'Integrative Genomics Viewer (IGV) (Robinson et al., 2011), un outil d'exploration en temps réel de multiples données génomiques ou transcriptomiques.

Le Sashimi plot est une représentation complexe pour l'œil de non-initiés et ne peut représenter qu'un nombre très limité d'événements d'épissage. Au-delà d'une dizaine d'évènements, la visualisation devient illisible. Ainsi, pour une représentation globale de la

qualité de l'épissage des introns, il est préférable de visualiser les distributions de l'ensemble des PSI associés à chaque rétention d'intron d'un échantillon à l'aide de boxplots (McGill et al., 1978) ou violinplots (Hintze and Nelson, 1998) (Figure 24). Ces méthodes de visualisation indiquent le premier, deuxième (médiane) et troisième quartile sous la forme d'une boîte (qui représente ainsi 50 % des données). Des traits verticaux au-dessous du premier quartile et au-dessus du troisième quartile, appelés les moustaches, peuvent avoir différentes significations selon la volonté de l'utilisateur. Le plus souvent, elles indiquent les intervalles entre [Q1 – 1.5 * IQR ; Q1] et [Q3 ; Q3 + 1.5 * IQR], avec Q1 et Q3 la valeur du premier et troisième quartile et IQR (InterQuartile Range) égale à Q3 – Q1. Si, dans les données représentées, il existe des valeurs extérieures aux limites de cette boîte à moustaches, elles seront affichées avec un point et considérées comme des points aberrants (outliers). Par rapport au box-plot, le violin-plot va en plus représenter la distribution de densité des valeurs observées.

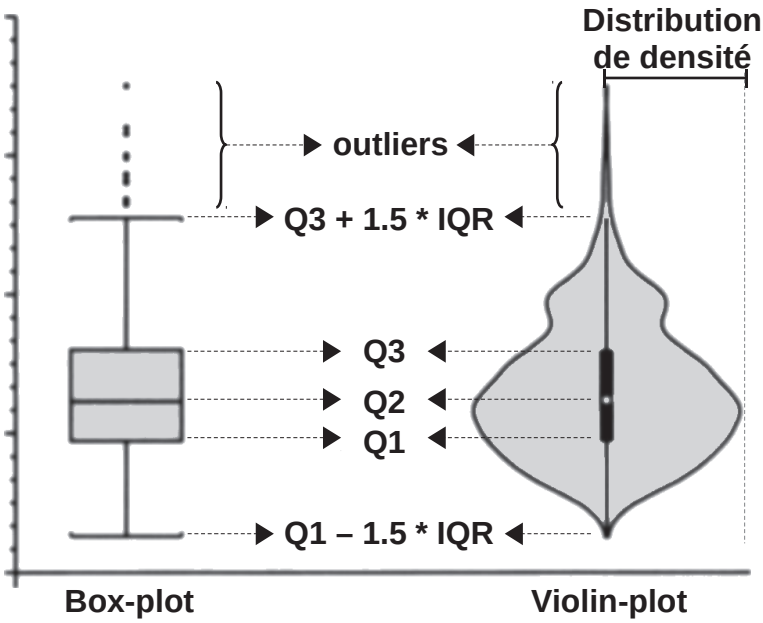

Figure 24 : Box- et violin-plots.
Présentation et comparaison des box-plot (ou boîte-à-moustaches) et violin-plot pour des données fictives. Q1, Q2 et Q3 représentent le premier, deuxième (ou médiane) et troisième quartile. IQR : InterQuartile Range (IQR = Q3-Q1). La distribution de densité des données est répétée des deux côtés du box-plot pour le violin-plot. *Source : (Hintze and Nelson, 1998).*

2 Objectifs

Le projet TALS est né en 2011 suite à l'identification des mutations de *RNU4ATAC*, qui est le résultat d'un travail collaboratif d'une équipe de généticiens et cliniciens (dirigée par Patrick EDERY) des hôpitaux de Bron avec une biostatisticienne de l'université de Paris (Anne-Louise LEUTENEGGER). Pour poursuivre les recherches sur ce syndrome, l'ARN de cellules provenant de fibroblastes (cellules de soutien du tissu conjonctif) de deux patients TALS, ainsi que des cellules de lymphoblastes (cellules lymphoïdes souches) pour l'un d'eux, ont été séquencés, avec des cellules d'individus contrôles (même tissu, âge et sexe). Une nouvelle collaboration avec des bioinformaticiens situés à Villeurbanne (Vincent LACROIX et Amandine FOURNIER puis Clara BENOIT-PILVEN) a permis de lancer une étude pilote sur l'analyse des résultats de ces séquençages. C'est dans ce contexte que s'est effectué mon stage de Master 2 bioinformatique entre 2015 et 2016, au Laboratoire de Biométrie et Biologie Évolutive (LBBE), sous la responsabilité de Vincent LACROIX, réalisé en collaboration avec l'équipe GENDEV (co-dirigée par Patrick EDERY et Sylvie MAZOYER) et Anne-Louise LEUTENEGGER. J'ai ainsi pu me familiariser avec ces équipes, l'épissage mineur, le syndrome de Taybi-Linder et les données de séquençage, ce qui m'a donné l'opportunité de poursuivre cette analyse transcriptomique en thèse, dans l'équipe GENDEV, et sur une cohorte de patients beaucoup plus grande, fruit d'une collecte minutieuse de 15 années. Le but de cette thèse est ainsi de caractériser les défauts transcriptomiques liés à un dysfonctionnement du spliceosome mineur, en utilisant le modèle du syndrome de Taybi-Linder.

Pour cela, les multiples domaines d'expertise trouvés au sein des équipes du projet TALS sont un atout capital. L'interaction continuelle entre bioinformaique, biostatistique et génétique, caractéristique du projet, pose un cadre de travail et d'entre-aide optimal pour ce sujet interdisciplinaire.

J'ai ainsi pu me focaliser sur la mise en place d'un pipeline bioinformatique, construit à partir de méthodes récentes, pour analyser globalement et différentiellement les profils d'épissage de cellules de patients et de contrôles, tout en prenant en compte les multiples tissus séquencés. Comme nous avons montré qu'une détérioration de la qualité de l'épissage pouvait induire un déficit d'expression, le profil d'expression de patients TALS devait aussi être caractérisé. Cette étude transcriptomique approfondie est la première portant sur des patients

atteints du syndrome de Taybi-Linder, et aussi celle composée du plus grand nombre d'individus et de tissus dans les cadres des maladies liées à un dysfonctionnement spécifique du spliceosome mineur.

Bien que notre intérêt premier fût d'étudier les cellules de patients, nous nous sommes également intéressés aux cellules contrôles, pour étudier l'épissage mineur dans des conditions normales. Cette analyse descriptive était d'autant plus importante que les rares études fondamentales sur l'épissage mineur ont été réalisées avant ou au début de l'avènement du séquençage massif des ARN, alors même que peu d'introns mineurs étaient caractérisés. Il était donc nécessaire de mettre à jour les connaissances scientifiques sur ce sujet, en proposant entre autres une nouvelle liste d'introns U12, pour poser les bases de l'étude comparative.

Cette thèse aspirait aussi à proposer des pistes entre épissage mineur et développement embryonnaire/cérébral, bien que la rareté des patients TALS complique énormément cette tâche. De plus, nous avons décrit la tissu-spécificité très étonnante des spliceosomopathies, et nous ne disposions d'aucun tissu cérébrale durant notre analyse. Nous savons aussi que le profil d'épissage des cellules est particulier dans le cerveau, et qu'il peut être extrêmement variable en fonction du stade de développement. Il est donc très difficile de s'avancer avec certitude sur le lien entre le spliceosome mineur et la formation ou le fonctionnement du cerveau.

3 Résultats

I. Pipeline d'analyse de l'épissage alternatif

Le pipeline KisSplice est une méthode originale d'analyse de l'épissage, puisqu'elle est la seule à utiliser l'assemblage des reads pour détecter localement les événements d'épissage. Comme KisSplice a été conçu au LBBE, laboratoire dans lequel s'est effectuée la moitié de ma thèse, j'ai pu le tester intensivement et participer au développement de briques logiciels de KisSplice, dont KisSplice2RefGenome (K2RG, écrit en python), pour lequel j'ai principalement réalisé de la détection/résolution de bugs et de l'optimisation des performances.

Dans sa publication, Clara BENOIT-PILVEN a comparé une méthode mapping-first (FaRLine) avec la méthode assembly-first KisSplice pour déterminer les forces et faiblesses de chacune d'elle dans le contexte de la détection et quantification de saut d'exons. En plus de ma contribution au développement de K2RG, utilisé dans cette publication, j'ai aussi activement participé aux discussions portant sur la comparaison des deux méthodes et à l'établissement de leurs spécificités.

Ce travail a été publié en Janvier 2018 dans la revue *Scientific Reports*. L'article et les figures supplémentaires sont inclus dans le paragraphe suivant. Les tableaux supplémentaires sont disponibles à l'adresse suivante : https://www.nature.com/articles/s41598-018-21770-7#Sec19.

A. Publication (*Scientific Reports*)

SCIENTIFIC REPORTS

OPEN

Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data

Received: 15 September 2017
Accepted: 30 January 2018
Published online: 09 March 2018

Clara Benoit-Pilven[1], Camille Marchet[3], Emilie Chautard[1,2], Leandro Lima[2], Marie-Pierre Lambert[1], Gustavo Sacomoto[2], Amandine Rey[1], Audric Cologne[2], Sophie Terrone[1], Louis Dulaurier[1], Jean-Baptiste Claude[1], Cyril F. Bourgeois[1], Didier Auboeuf[1] & Vincent Lacroix[2]

Genome-wide analyses estimate that more than 90% of multi exonic human genes produce at least two transcripts through alternative splicing (AS). Various bioinformatics methods are available to analyze AS from RNAseq data. Most methods start by mapping the reads to an annotated reference genome, but some start by a *de novo* assembly of the reads. In this paper, we present a systematic comparison of a mapping-first approach (FARLINE) and an assembly-first approach (KISSPLICE). We applied these methods to two independent RNAseq datasets and found that the predictions of the two pipelines overlapped (70% of exon skipping events were common), but with noticeable differences. The assembly-first approach allowed to find more novel variants, including novel unannotated exons and splice sites. It also predicted AS in recently duplicated genes. The mapping-first approach allowed to find more lowly expressed splicing variants, and splice variants overlapping repeats. This work demonstrates that annotating AS with a single approach leads to missing out a large number of candidates, many of which are differentially regulated across conditions and can be validated experimentally. We therefore advocate for the combined use of both mapping-first and assembly-first approaches for the annotation and differential analysis of AS from RNAseq datasets.

In the last 10 years, the prevalence of alternative splicing has been completely re-evaluated. Recent reports claim that more than 90% of multi-exon genes produce at least two splicing variants[1,2]. The depth at which transcriptomes can be sampled with next generation sequencing techniques opens the possibility not only to annotate splicing variants in various conditions, but also to detect which transcripts are differentially spliced across pathological and physiological conditions.

This growing interest in splicing both as a fundamental process and because of its implication in pathologies[3–5] has been accompanied by an increasing number of methods aiming at analyzing RNAseq datasets[6–8]. The ultimate goal of these methods is to identify and quantify full-length transcripts from short sequencing reads. This task is particularly challenging and recent benchmarks show that all methods still make a lot of mistakes[9]. The difficulty of reconstructing full-length transcripts (isoform-centric approaches) also prompted a number of authors to focus on identifying exons that are differentially included within transcripts (exon-centric approaches)[10–13].

Whether they are exon- or isoform-centric, methods to study splicing from RNAseq data can further be divided in two main categories[14]. The mapping-first approaches first map the reads to the reference genome and

[1]Université de Lyon, ENS de Lyon, Université Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France. [2]Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, EPI ERABLE - Inria Grenoble, Rhône-Alpes, France. [3]IRISA Inria Rennes Bretagne Atlantique CNRS UMR 6074, Université Rennes 1, GenScale team, Rennes, 263 Avenue Général Leclerc, Rennes, France. Correspondence and requests for materials should be addressed to D.A. (email: Didier.auboeuf@inserm.fr) or V.L. (email: Vincent.lacroix@univ-lyon1.fr)

the mapped reads are then assembled into exons and eventually transcripts. In contrast, assembly-first approaches first assemble the reads based on their overlaps. The assembled sequences (corresponding to sets of exons) are then aligned to the reference genome.

Mapping-first approaches have been the most used so far, essentially because they were the first to be developed and because they initially required less computational resources. *De novo* assembly methods were also thought to be restricted to non-model species, where no (good) reference genome is available, and they seemed to be inadequate when an annotated reference genome is available.

Recent progress in *de novo* transcriptome assembly is clearly changing this view, and the argument of the heavier computational burden does not hold anymore.

The application of *de novo* assembly to human RNAseq datasets however still remains rare, although some studies have already shown its potential to detect novel biologically relevant splicing variants[15,16].

The generalization of *de novo* assembly approaches for studying splicing in human seems to be mostly impeded by the lack of a clear evaluation of its potential interest in comparison to more traditional mapping-based approaches.

This is the gap we aim at filling with the work presented here.

To achieve this goal, we performed a systematic evaluation of an assembly-first and a mapping-first approach on two RNAseq datasets.

As a first step, we compared pipelines that we developed in parallel, namely KisSplice and FaRLine, because we could easily control their parameters. Any difference between the predictions that is solely due to a parameter setting could be fixed easily, which enabled us to obtain a precise understanding of the irreducible differences between the two approaches.

In a second step, we confirmed the generality of our findings by benchmarking our methods against Cufflinks[6], MISO[11] and Trinity[17], which are widely used pipelines.

A significant part of our work has been to manually dissect a number of cases found by only one of the two methods. This enabled us to go beyond a simple qualitative description and provide the community with a precise understanding of which cases are overlooked by each type of method, and where new methods are needed.

All the software and step-by-step protocols presented in this work are freely available at http://kissplice.prabi.fr/pipeline_ks_farline. This should facilitate the reproducibility of our work, and applications to other datasets.

From a general point of view, the combination of approaches we propose should enable to improve splicing-related transcriptomic analyses in physiological and pathological situations.

Results

KisSplice and FaRLine. Figure 1 presents schematically the two pipelines that we developed and compared. A detailed description of each step is given in the Methods section. In the assembly-first approach, a De Bruijn graph is built from the reads. Alternative splicing events, which correspond to bubbles in this graph are enumerated and quantified by KisSplice. Each path is then mapped on the reference genome using STAR and the event is annotated by KisSplice2RefGenome, using the EnsEMBL r75 annotations as an evidence. Importantly, exons not present in the annotations can be identified by this approach. In the mapping-first approach, reads are aligned to the reference genome using TopHat2. Mapped reads are then analyzed by FaRLine, using the EnsEMBL r75 annotations as a guide.

We also tested STAR instead of TopHat2 for the mapping-first pipeline, and found that our main results were essentially unchanged (see Methods).

Quantification of splicing variation is performed similarly in the two pipelines. Only junction reads are considered. Exonic reads are not considered, for reasons exposed in Methods. For the inclusion isoform, there are two junctions to consider. We calculate the mean of the counts of these two junctions.

The differential analysis is performed by a common method for the two approaches: kissDE, which tests if the relative abundance of the inclusion isoform has changed significantly across conditions.

Overall, we developed and adapted jointly these two pipelines in order to minimize the discrepancies that could complicate the comparison.

The majority of frequent isoforms are identified by both approaches. Applying KisSplice and FaRLine to the same RNAseq datasets generated by the ENCODE consortium (SK-N-SH cell lines treated or not with retinoic acid), we noticed that 68% of the alternatively skipped exons (ASE) identified by KisSplice were also identified by FaRLine and that 24% of ASEs identified by FaRLine were also identified by KisSplice (Fig. 2A). This observation highlights that the mapping-first approach predicts a much larger number of events. This difference in sensitivity is due to the fact that while mapping-first approaches require that each exon junction is covered by at least one read, assembly-first approaches require overlapping reads across the entire skipped exon. Therefore, it can be anticipated that low abundant isoforms, that are covered by few reads, will be reported by mapping, but not by the assembly-first approach. Supporting this prediction, we observed that for ASEs reported only by FaRLine, the number of reads supporting the minor isoform is much lower than in the other categories (Fig. 2 B). The same results were obtained using another RNAseq dataset representing MCF-7 cells expressing or not the DDX5 and DDX17 splicing factors (Supplementary Figure S1).

Having clarified that rare variants are better handled by the mapping-first approach, we decided to filter them out, in order to analyse other differences between the two approaches. Experimental validations by RT-PCR that we performed on rare variants stratified by read support enabled us to clarify that both an absolute and a relative cutoff on the number of reads are required to discriminate variants which can be validated from those which cannot. Indeed, out of the 48 tested cases, we were able to validate 41 (Supplementary Figure S9). The non validated cases indeed corresponded to cases supported by fewer reads. However, what really departed them from the validated cases was their lower relative abundance (Supplementary Figure S10, Supplementary Table 1). In the

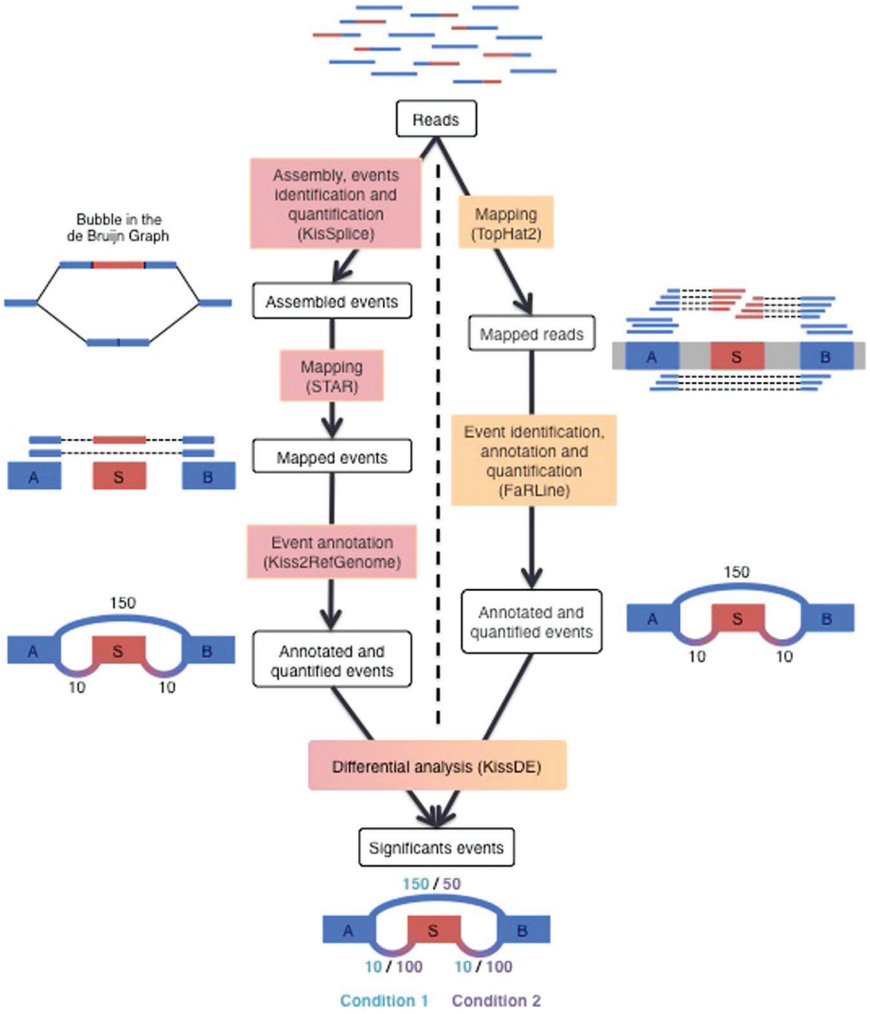

Figure 1. The two pipelines compared in this study: KisSplice and FaRLine. The first step of KisSplice is to assemble the reads and extract the splicing events. These events are then mapped back to the reference genome and classified by event type. The annotated and quantified events are then used for the differential analysis between the biological conditions. In contrast, the first step of FaRLine is to map the reads on the reference genome. From this mapping, annotated and quantified events are extracted. Finally, the differential analysis is done with the same method as in the KisSplice pipeline.

remaining of our work, we chose to use both criteria and we filtered variants supported by less than 5 reads, and less than 10% compared to the major isoform.

As expected, the proportion of candidates reported simultaneously by both methods increased significantly. Approximately 70% of predicted skipped exons were indeed found by both approaches after filtering lowly expressed isoforms. (Fig. 2C, Supplementary Figure S1C).

Furthermore, the estimation of their inclusion rates was consistent across the two approaches ($R^2 > 0.9$)).

Beyond the overall concordance of the two approaches in detecting common splicing events, a number of candidates remained reported by only one approach. Since many of them have a highly-expressed minor isoform (supported by more than 100 reads) (Fig. 2D, Supplementary S1D), the failure of one approach to detect them is likely not due to a lack of coverage.

For events only found by one approach, we patiently dissected the reasons why they could have been missed out by the other approach. This enabled us to define 4 main categories which cover 70% of the cases (Fig. 3A) The remaining 30% of cases did not fit into clearly defined biological categories. We however classified them using methodological criteria. The full list of categories is presented in Supplementary Table 2. For each of the 4 main categories, we selected cases to validate experimentally. All 34 RT-PCR validations were successful and are presented in Supplementary Figure S11 confirming that these events are not false positives.

Some isoforms are systematically missed by one approach. The first category corresponds to cases that were missed out by the mapping-first approach and corresponds to alternative splicing events involving novel exons or novel combinations of existing exons.

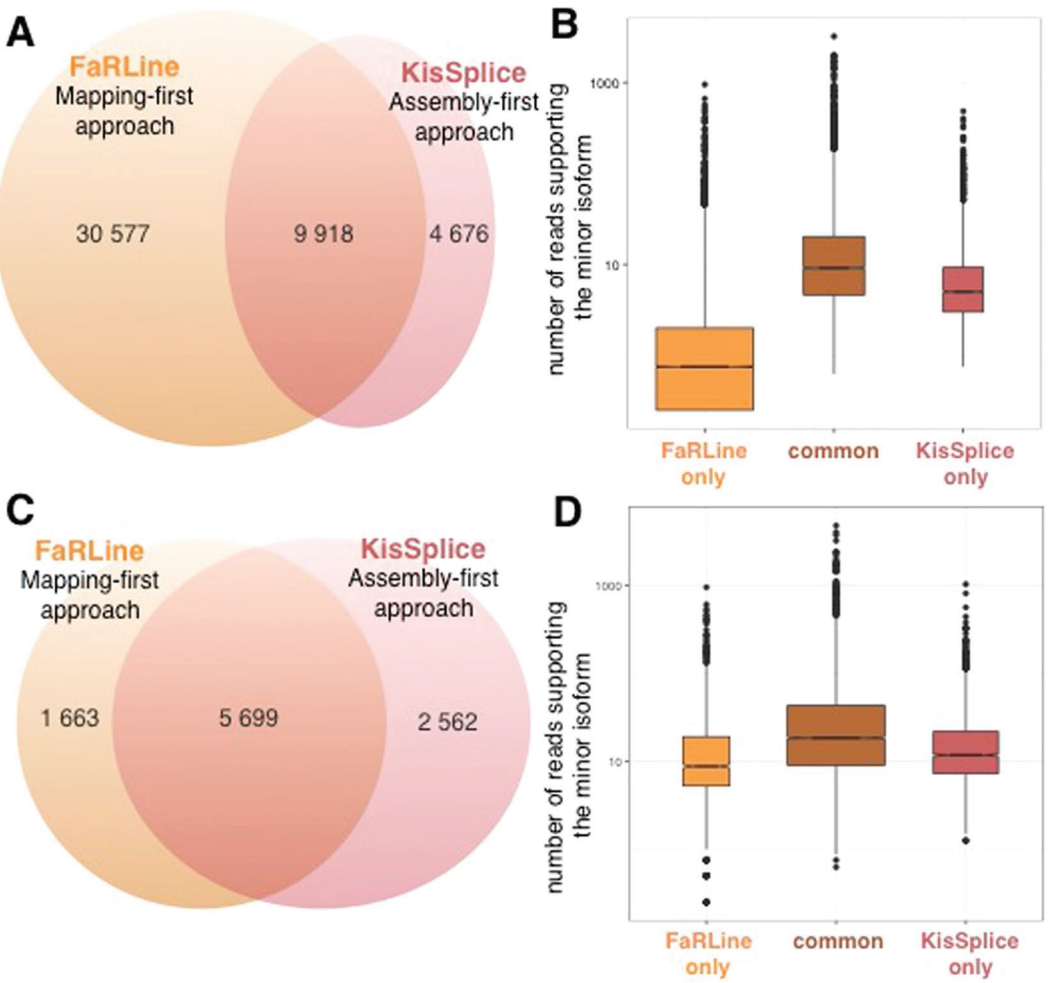

Figure 2. Comparison of the ASE identified by the assembly-first and mapping-first pipelines. (**A**) Venn diagram of ASEs identified by the two pipelines. FaRLine detected many more events than KisSplice. 68% of ASE found by KisSplice were also found by FaRLine and 24% of ASE detected by FaRLine were also found by KisSplice. (**B**) Boxplot of the expression of the minor isoform in the 3 categories defined in the Venn diagram of panel A: ASE identified only by FaRLine, ASE identified by both pipelines and ASE identified only by KisSplice. The number of reads supporting the minor isoform of the ASE identified by FaRLine is overall much lower. Many isoforms are supported by less than 5 reads. (**C**) Venn diagram of ASEs identified by the two pipelines after filtering out the poorly expressed isoforms (less than 5 reads, or less than 10% of the number of reads supporting both isoforms). The common events represent a larger proportion than before filtering: 77% of the ASE identified by FaRLine and 69% of the ASE identified by KisSplice. (**D**) Boxplot of the expression of the minor isoform in the 3 categories defined in the Venn diagram of panel C: ASE identified only by FaRLine, ASE identified by both pipelines and ASE identified only by KisSplice. The distribution of the number of reads supporting the minor isoform is similar for the 3 categories with highly expressed variants in each category.

There are two reasons to explain why the mapping-first approach does not detect these events. First the mapper may fail to map the reads, or map them to an incorrect location, as junction discovery using short reads is a challenging task. Second, even in the case where the mapper succeeds, FaRLine may fail to report the event because it relies on annotations. Among these 1864 cases, we distinguished 3 sub-categories of errors due to the annotation. Either the exon is unannotated (30%), one of its flanking exon is unannotated (13%) or both exons are annotated but no transcript combining them was annotated (57%).

The assembly-first approach, KisSplice, does not consider annotations, and an interesting resulting advantage is that novel junctions have the same chance to be assembled as known junctions. Mapping assembled novel junctions to the genome is indeed less challenging than read mapping because the assembled sequences are longer.

More importantly, the ability of KisSplice to identify novel splicing events comes from the fact that it introduces known annotations as late as possible in its pipeline (see Methods). Annotations are used as an evidence, not as a filter. AS events involving novel splice sites are clearly identified as such, and can be specifically tested and experimentally validated. More than 99% of the novel splice sites were canonical splice sites (GT-AG).

As an example, the *HIRA* gene contains a novel exon, whose inclusion is supported by at least 20 reads on each junction (Fig. 3B, Supplementary Figure S8A). This case was overseen by the mapping-first approach, FaRLine.

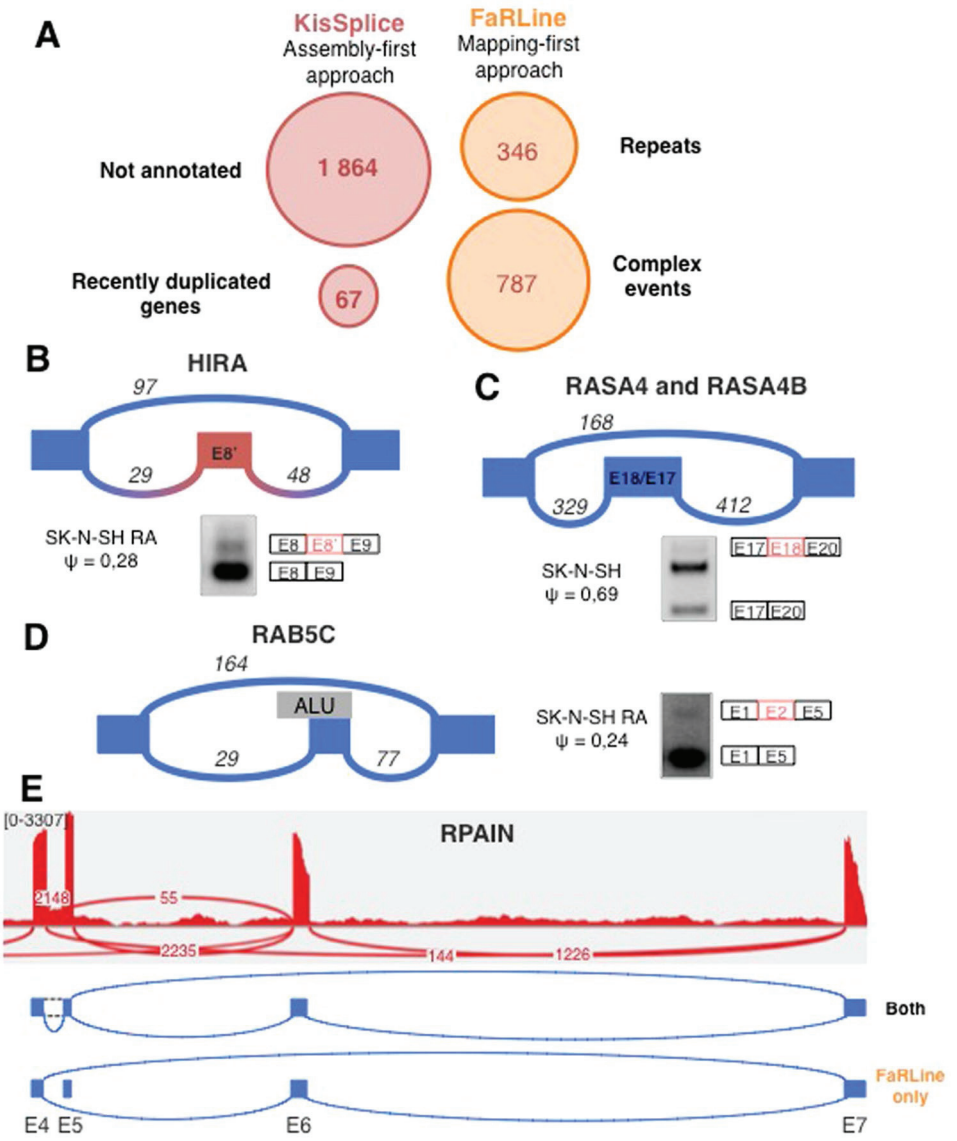

Figure 3. (**A**) Main categories explaining why some exons are detected by only one method. (**B**) The exon in intron 8 of the *HIRA* gene is an example of an exon not annotated in EnsEMBL r75. This event was identified by KisSplice but not by FaRLine. (**C**) *RASA4* and *RASA4B* are 2 paralog genes. KisSplice detected 2 isoforms that could be produced by these 2 genes. FaRLine did not detect any event in either of these genes. The exon skipped is exon 18 in *RASA4* (corresponding to exon 17 in *RASA4B*). The third band on the RT-PCR is the inclusion of another exon in the intron 18 of *RASA4*. (**D**) Exon 2 of the *RAB5C* gene is an example of exon skipping overlapping an Alu element identified only by FaRLine. The events in panel B to C were validated by RT-PCR. (**E**) The *RPAIN* gene contains a complex event with a lowly expressed isoform. This weakly expressed isoform was not identified by KisSplice, while the other isoforms were identified by both approaches.

The panel B of the Supplementary Figure S8 shows an example of an ASE not reported by FaRLine because the included exon was not present in the transcripts.

The second category of splicing events identified by only one approach corresponds to recent gene duplications. Untangling the relation between alternative splicing and gene duplication is a difficult topic, subject to debate[18,19]. It is indeed difficult to assess the amount of alternative splicing that occurs within paralogous genes. With the mapping-first approach, the reads stemming from recent paralogs are classified as multi-mapping reads. FaRLine, like the vast majority of mapping-first pipelines, discards these reads for further analysis, as their precise location cannot be clearly established. This results in silently underestimating alternative splicing in recent paralog genes. Note that setting the mapper to keep multi-mapping reads in the analysis leads to overestimating alternative splicing, as all members of the family will be predicted as alternatively spliced. In opposition, *de novo* assembly can faithfully state that a family of recent paralogs collectively produce two isoforms that vary in their sequence. However, whether the two isoforms are produced from the same locus or from different loci remains undetermined. KisSplice detects these cases of putative AS in paralog genes. Figure 3C illustrates the case with

genes *RASA4* and *RASA4B*. Exon 18 in *RASA4* (denoted as exon 17 in *RASA4B*) was detected to be skipped. The exclusion isoform is supported by 160 reads, while the inclusion isoform is supported by 400 reads. The mapping-first approach did not detect either of these isoforms at all. Another example from this category is presented in Supplementary Figure S2C.

The third category of splicing events identified by only one approach corresponds to cases that are missed out by the assembly-first approach. Out of the 1663 cases belonging to this category, a large fraction (21%) corresponds to cases where the skipped exon overlaps a repeat, notably Alu elements. Alu are transposable elements present in a very large number of copies in the human genome[20]. Most of these copies are located in introns and a number of them have been exonised[21,22]. The reason why the mapping-first approach is able to identify these cases is because even though the reads partially map to repeated sequences, the boundaries of these exons are unique and annotated. Hence the mapper, if set properly, can map these reads to unique annotated exon junctions and is not confused by multiple mappings. Importantly, if the annotations are not provided to the mapper, it will be confused by multiple mappings and will not be able to map the read to the correct location (Supplementary Figure S7). Figure 3D and Supplementary Figure S2D represent two RT-PCR validated Alu-derived exons identified by the mapping-first approach. The assembly-based approach fails to detect most of these events. The reason is that, although they do form bubbles in the DBG generated by the reads, these bubbles are highly branching (supplementary figure http://kissplice.prabi.fr/sknsh/graph_RAB5C_distance_3.html23). Enumerating branching bubbles is computationally very challenging, and may take a prohibitive amount of time. In practice, we restrict our search to the enumeration of bubbles with at most 5 branches (Supplementary Figure S12A).

The fourth category of splicing events identified by only one approach corresponds to cases where more than two splicing isoforms locally coexist, and one of them is poorly expressed compared to the others. The *RPAIN* gene is a good illustration of such cases (Fig. 3E), as exons 5 and 6 of *RPAIN* may be skipped and the intron between exons 4 and 5 may be retained. While both methods successfully reported the skipping of exon 6, with exons 5 and 7 as flanking, FaRLine additionally reported the skipping of the same exon, but with exons 4 and 7 as flanking exons. The reason why KisSplice did not report this case is because the junction between exons 4 and 6 is relatively weakly supported. More specifically, this junction is supported by only 55 reads, which accounts for less than 2% of the total number of reads branching out from exon 4. Transcriptome assemblers, like KisSplice, usually interpret such relatively weakly supported junctions as sequencing errors or spurious junctions in highly-expressed genes, therefore disregarding them in the assembly phase (see Supplementary Methods). Supplementary Figure S2E shows another example of a complex event not correctly handled by KisSplice because there were locally more than 5 branches.

Comparison of the approaches after differential analysis. Beyond the tasks of identifying exon skipping events, a natural question which arises when two conditions are compared is to assess if the exon inclusion rate significantly changed across conditions.

In order to test this, we took advantage of the availability of replicates for both the SK-N-SH cell line and the same cell line treated with retinoic acid. For each detected event, we tested with kissDE[24], whether we could detect a significant association between one isoform and one condition. Focusing on those condition-specific events, we again partitioned them in events reported by both methods, by FaRLine only and by KisSplice only. As shown in Fig. 4, the majority of condition-specific events were detected by both approaches. This is the case for instance of exon 22 of gene *ADD3* which is clearly more included upon retinoic acid treatment (Fig. 4C), with a DeltaPSI of 27%. The estimation of the DeltaPSI is overall very similar across the two approaches (Fig. 4B) with a correlation of 0.94. The outliers essentially correspond to ASE with several alternative donor/acceptor sites. KisSplice considers these events as different exons while FaRLine considers them as an unique exon, and sums up all the incoming (resp. outgoing) junction counts. Hence, the read counts will differ. Supplementary Figure S8D gives an example.

When focusing on condition-specific events, the proportion of events predicted by only one method increased, for two main reasons. First, some ASE annotated by both approaches were predicted to be differentially included only by one method. This is again due to differences in the quantification of the inclusion rate, especially for ASE with multiple 5′ and 3′ splice sites. Second, some of the exons that were missed out by one method at the identification step happened to be condition specific. This is the case of an exon in *NINL* intron 5 (Fig. 4D), only identified by KisSplice because it was not annotated. This is also the case of *SAR1B* exon 3 (Fig. 4E), only identified by FaRLine because it overlaps with an Alu element. The analysis of the MCF-7 RNAseq dataset gave very similar results (Supplementary Figure S3).

The observation that many of the AS events that were annotated only by one method are differentially regulated across conditions confirms that these AS events should not be discarded from the analysis. Focusing only on AS events annotated by one approach may lead to miss splicing events which are central in the biological context.

Overlap with other methods. In a first step, we picked FaRLine and KisSplice as examples of a mapping-first and an assembly-first approach respectively. Clearly, there are other published methods in both categories. MISO is probably the most widely used to annotate AS events. We therefore ran it on the same datasets to check how its predictions overlapped with ours. As shown in Fig. 5A (SK-N-SH dataset), 77% of predictions made by MISO were common to both FaRLine and KisSplice, 18% were only common with FaRLine, 2% were only common to KisSplice and the remaining 3% were specific to MISO. The overlap between the different methods was very similar when the MCF-7 RNAseq dataset was used (Supplementary Figure S4A). Overall, almost all candidates predicted by MISO were also predicted by FaRLine. This large overlap with FaRLine was expected, because both are mapping-first approaches. This also shows that the differences between mapping- and assembly-first approaches reported above are not limited to one mapping-first approach.

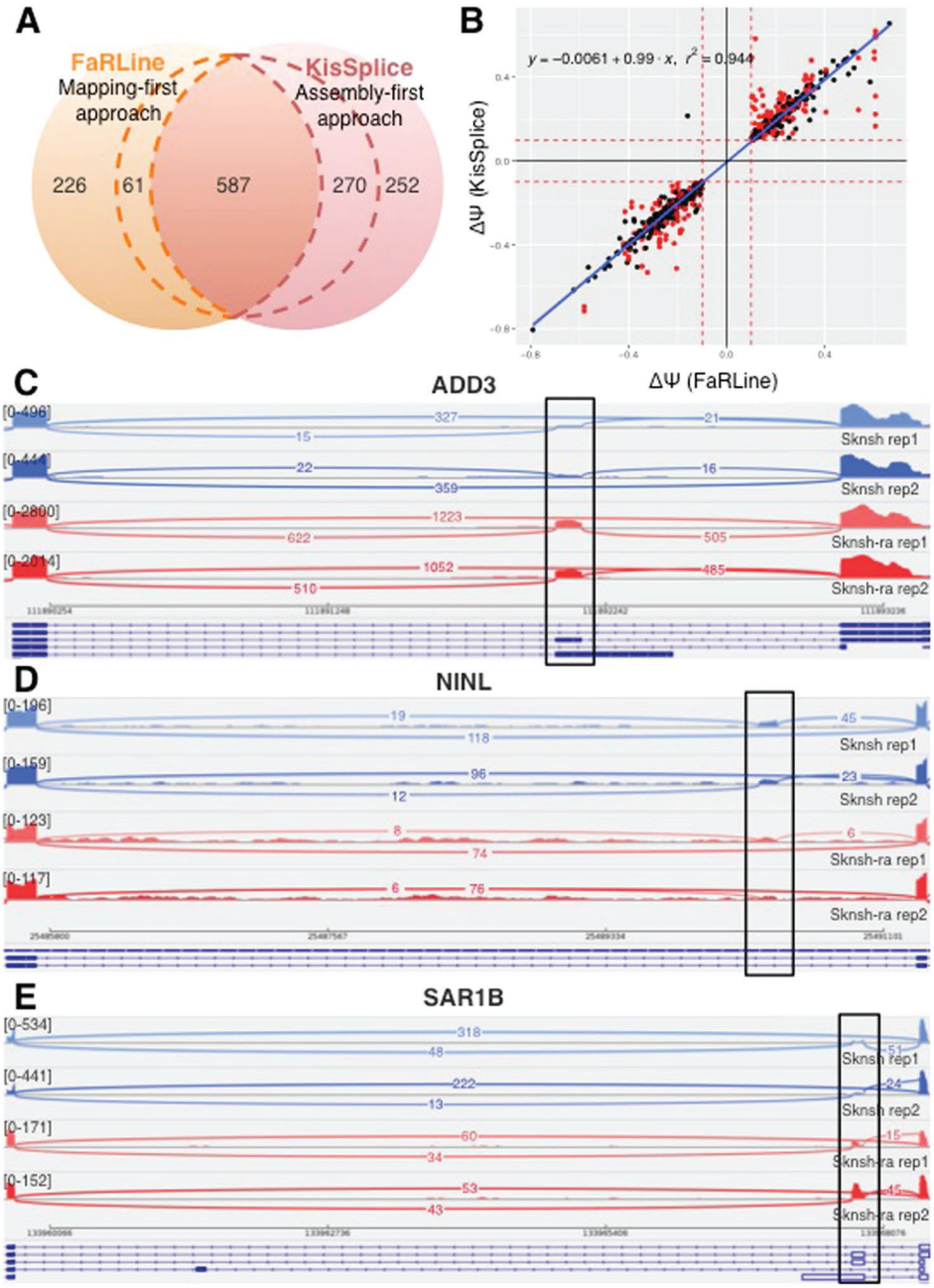

Figure 4. (**A**) Condition-specific variants identified by FaRLine, KisSplice or both methods. Within dashed lines are events identified by both approaches but detected as condition-specific by only one approach. (**B**) DeltaPSI as estimated by KisSplice and FaRLine, for events identified by both methods. The red dots represent complex events for which KisSplice found at least 2 'bubbles'. (**C**) Exon 22 of the *ADD3* gene is an example of regulated ASE identified by both approaches. (**D**) A new exon in intron 5 of *NINL* gene is identified only by KisSplice. The inclusion of this exon is differentially regulated between the 2 experimental conditions. (**E**) Because exon 3 of the *SAR1B* gene is an exonised Alu element, only FaRLine identified this event. Moreover this exon is significantly more included in the treated cells (SK-N-SH RA) compared to the control cells.

Besides exon-centric approaches, which aim at finding the differentially spliced exons, there is also a number of published methods which are isoform-centric and have the more ambitious goal to reconstruct full-length transcripts at the expense of underestimating alternative splicing.

The most widely used mapping-first and isoform-centric approach is Cufflinks[6] that we compared to FaRLine using the same dataset. As shown in Fig. 5B (and Supplementary Figure S4B), we found that the vast majority of ASE were predicted by both approaches.

Finally, we compared KisSplice to one of the most widely used de-novo transcriptome assembler, Trinity[17]. As shown in Fig. 5D (and Supplementary Figure S4D), most ASE found by Trinity were also found by KisSplice. However, KisSplice was significantly more sensitive. The goal of Trinity is to assemble the major isoforms

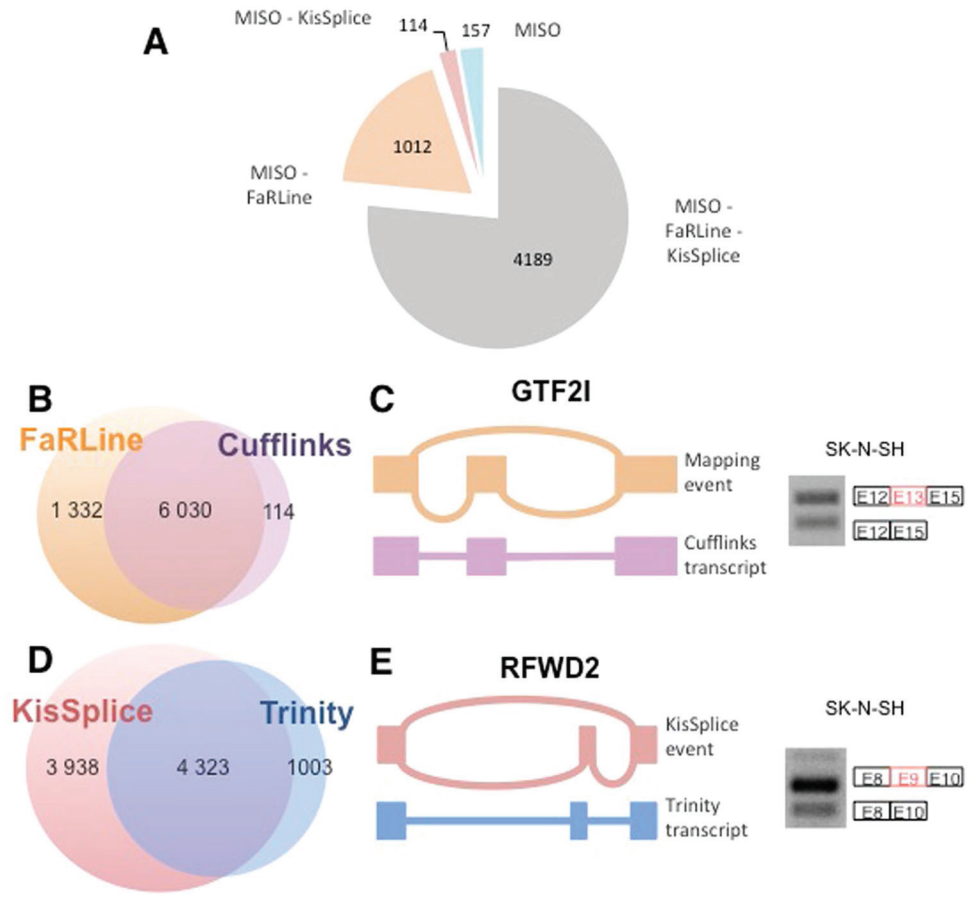

Figure 5. (**A**) 77% of ASE identifed by MISO are also annotated by FaRLine and KisSplice. 18% of MISO's ASE are also annotated by FaRLine while only 2% of MISO's ASE are also annotated by KisSplice. Finally, only 3% of these ASEs are only annotated by MISO. (**B**) Most of the events annotated by Cufflinks are identified by FaRLine. (**C**) *GTF2I* exon 13 is an example of an ASE annotated by FaRLine but not by Cufflinks. Indeed, Cufflinks only identified the isoform corresponding to the exon inclusion. (**D**) Most of the events annotated by Trinity are also annotated by KisSplice. But half of the ASE annotated by KisSplice are not annotated by the global assembler Trinity. (**E**) KisSplice annotates an ASE in the *RFWD2* gene, while Trinity only identified the isoform corresponding to the exon inclusion. The events in panels C and E have been validated by RT-PCR.

for each gene, it therefore largely under-estimates alternative splicing, especially inclusion/exclusion of short sequences.

For completeness sake, we also provide an all-vs-all comparison (Supplementary Figure S5). An interactive version of this Figure is available at http://kissplice.prabi.fr/pipeline_ks_farline/. The list of events found by any used method can be retrieved from this interactive figure and analysed in IGV, to reproduce the sashimi plots of the paper. The general conclusions from these comparisons is that there is a clear distinction between mapping-first and assembly-first approaches, and between exon-centric and isoform-centric approaches, the latter being less sensitive.

Discussion

De novo assembly is usually applied to non-model species where no (good) reference genome is available. We show here that even when an annotated reference genome is available, using assembly offers a number of advantages. We named this approach "assembly-first" because it does use a reference genome, but as late as possible in the process, in order to minimize the *a priori* on which exons should be identified.

Using this strategy, we identified novel alternatively skipped exons, which were not identified by traditional read mapping approaches (Fig. 3 and Supplementary Figure S2). While it is believed that the human genome is fully annotated, it is important to underline that we have not yet established a final map of the parts of the genome that can be expressed. It can be anticipated that sequencing of single-cells from different parts of the body will lead to the discovery of a huge diversity of transcripts and that a substantial number of new exons will be discovered. An example is the case of unannotated skipped exons which overlap with repeat elements. We cannot exclude that this category is currently largely under-annotated.

We also showed that assembly-first approach has the ability to detect splicing variants within recently duplicated genes (Fig. 3 and Supplementary Figure S2). This is because mapping approaches discard reads which map to multiple genomic locations. Identification of such splicing variants produced from different genomic regions

sharing sequence similarities (e.g. paralog genes, pseudogenes) is however very important, since splicing variants generated from paralogous genes but also from pseudogenes may have different biological functions[25].

Conversely, we showed that some ASE were detected only by the mapping-first approach. As shown in Fig. 2 (and Supplementary Figure S1), we observed that the mapping-first approach has a better ability to detect lowly-expressed splicing variants. Although such lowly-expressed splicing variants are often considered as "noise" or biologically non relevant, caution must be taken with such assumptions for several reasons. First, mRNA expression level is not necessarily correlated with protein expression level. Second, as observed from single-cell transcriptome analyses, some mRNAs can be expressed in few cells, within a cell population (e.g. they are expressed at a specific cell cycle step) and may therefore appear to be expressed at a low level in total RNAs extracted from a mixed cell population[26]. Therefore, computational analysis should not systematically discard lowly-expressed splicing variants and filtering these events should depend on the biological questions to be addressed.

We also observed that the mapping-first approach better detects exons corresponding to annotated-repeat elements (Fig. 3 and Supplementary Figure S2). While it has been assumed for a long time that repeat elements are "junk", increasing evidences support important biological functions for such elements. For example, repeat elements like Alu can evolve as exons and the presence of Alu exons in transcripts has been shown to play important regulatory functions[22,27].

When two methods give non-overlapping predictions, the temptation could be to focus on exons found by both approaches and to discard the others. We argue that this is not the best option, because approach-specific cases can be validated experimentally, and also because many of them correspond to regulated events, i.e. the inclusion isoform is significantly up or down regulated depending on the experimental condition.

In conclusion, combining mapping- and assembly-first approaches allows to detect a larger diversity of splicing variants. This is very important towards the in depth characterization of cellular transcriptome although other approaches are further required to analyze their biological functions.

From a computational perspective, a number of challenges are still ahead. The co-development of two approaches enabled us to narrow down the list of difficult instances not properly dealt with by at least one approach, but we cannot exclude that some categories are still missed out by both approaches. The categories of challenging cases that we defined in Fig. 3: lowly-expressed variants, exonised Alu, complex splicing variants, paralogs have been overlooked up to now. Possibly because they are much harder to detect, they have been assumed to play a minor role in transcriptomes, but more recent studies however argues the opposite.

For exonised ALUs, paralog genes and genes with complex splicing patterns, the possibility to sequence longer reads with third generation techniques[28,29] should prove very helpful. The number of reads obtained with these techniques is however currently much lower than with Illumina, thereby preventing their widespread use for differential splicing, for which the sequencing depth, and not so much the length of the reads, is the critical parameter which conditions the statistical power of the tests. In the coming years, methods combining second and third generation sequencing should enable to obtain significant advances in RNA splicing.

Material and Methods

FaRLine and KisSplice. Figure 1 shows the two pipelines that we are comparing. While STAR and TopHat are third-party softwares, we developed the other methods ourselves. KisSplice was first introduced in Sacomoto *et al.*[13]. The novelty here is that its usage is now possible in the case where a reference genome is available, which required specific methodological developments implemented in the newly released KisSplice2RefGenome software. kissDE was first introduced in Lopez-Maestre *et al.*[24] in the context of SNPs for non-model species. We present here its extension for alternative splicing. FaRLine is a new mapping-first pipeline, that we introduce in this paper. It is the RNAseq pipeline associated to the FasterDB database[30] and was already successfully applied to the analysis of the effect of metformin treatment on myotonic dystrophy type I (DM1) with a validation rate of 95%[31]. Specifically, 20 cases of ASE regulated by the metformin treatment were tested, and 19 were validated. In this paper, we provide additional validations of FaRLine with similar validation rates (36 out of 38), Supplementary Figure S19.

For the sake of self-containment, we explain all methods here.

KisSplice. KisSplice is a local transcriptome assembler. As most short reads transcriptome assemblers[8,17,32], it relies on a De Bruijn graph (DBG). Its originality lies in the fact that it does not try to assemble full-length transcripts. Instead, it assembles the parts of the transcripts where there is a variation in the exon content. By aiming at a simpler goal, it can afford to be more exhaustive and identify more splicing events. The key concept on which KisSplice is built is that variations in the nucleotide content of the transcripts will correspond to specific patterns in the DBG called bubbles (Supplementary Figure S13). KisSplice's main algorithmic step therefore consists in enumerating all the bubbles in the graph built from the reads. Examples of bubbles in the DBG and explanation of the parameters used to filter out sequencing errors and repeat-induced bubbles are given in Supplementary Methods.

Annotating the events with KisSplice2RefGenome. KisSplice outputs bubbles in the form of a pair of fasta sequences. Clearly, such information is insufficient to analyse alternative splicing for model species. KisSplice2RefGenome enables to provide for each bubble: the gene name, the AS event type, the genomic coordinates and the list of splice sites used (novel or annotated).

Bubbles found by KisSplice are mapped to the reference genome using STAR, with its default settings, which means that in the case of multi-mappings, STAR reports all equally best matches. The mapping results are then analysed by KisSplice2RefGenome. Bubbles are classified in sub-types depending on the number of blocks obtained when mapping each path of the bubble to the genome (Supplementary Figure S14). For exon skipping,

the longer path of the bubble corresponds to 3 blocks, while the lower path corresponds to 2 blocks. The splice sites are located and compared to the annotations. Events with novel splice sites are reported explicitly as such in the output of the program.

In the case where the bubble corresponds to a genomic insertion or deletion, it exhibits a specific pattern in terms of block numbers (one block for one path and two blocks for the other) and is reported separately.

The criterion of the number of blocks is discriminative in most cases. However, there is a possible confusion between intron retentions and genomic deletions, since in both cases, the longer path will map into one block and the lower path in two blocks. In this case, we also use the distance between the blocks, and introduce a user-defined threshold, which we set to 50nt, below which the bubble is classified as a genomic deletion, and above which it is classified as an intron retention.

In the special case where the exon flanking the AS event is very short (less than k nt), the number of blocks is increased for both paths, but the difference of number of blocks remains unchanged.

In the special case where there is a genomic polymorphism located less than k nt apart from the AS event, KisSplice will report several bubbles (possibly all combinations of genomic and transcriptomic variants). This redundancy is removed in KisSplice2RefGenome where the primary focus in on splicing.

In the case where the bubble maps to two locations on the genome, a distinction is made between the case of exact repeats where both paths map to both locations and inexact repeats where each path maps to a distinct location (Supplementary Figure S12B). The cases of exact repeats correspond to recent gene duplications.

FaRLine. **FasterDB EnsEMBL r75 annotation:** FasterDB RNAseq Pipeline, FaRLine, uses the FasterDB-based EnsEMBL r75 annotation database. FasterDB is a database containing all annotated human splicing variants[30].

Each transcripts present in the FasterDB, is composed of a succession of exons, that we call transcript exons (represented in blue in Supplementary Figure S15). The genomic exons (represented in red in Supplementary Figure S15) are defined by projecting the transcript exons. First, the transcript exons are grouped by position. Then each group of exons defines a projected exon with the following rules:

- The start is the leftmost start of the non-first-exon of the group.
- The end is the rightmost end of the non-last-exon of the group that ends before the start of the next group of exons.

When the most frequent event annotated in the transcripts is an intron retention, the projected genomic exon is defined as a combination of the two exons flanking the retained intron. In Supplementary Figure S15, the exons 5 and 6 and the intron 5 are considered as one unique exon. As events included within one exon are not tested, this results in some events being missed.

Mapping: The first step of FaRLine is to map the reads to a reference genome. This step is done using Tophat-2.0.11[6]. tophat–min-intron-length 30–max-intron-length 1200000\-p 8 [–solexa1.3-quals for Sknsh_rep1 and Sknsh_rep2]\–transcriptome-index

A transcriptome index has been built by TopHat using EnsEMBL r75 annotations in gtf format. When a transcriptome index is used, the mapping steps are modified: instead of aligning first to the genome, which is the default behavior, TopHat uses Bowtie to align the reads to the transcript sequences first, then align the remaining unmapped reads to the genome. Minimal and maximal intron lengths have been modified (default 70 and 500000) to maximize the number of junctions detected, according to the statistics provided by FasterDB EnsEMBL r75 annotations.

The resulting alignment files have been filtered using samtools 0.1.19[33].

Samtools view -F 260 -f 1 -q 10 -b

With this step, only the primary alignments are kept. The minimum read alignment quality was set up so that multi-mapping reads were removed from the alignment file.

Annotation and quantification of alternative splicing events: For each gene, all the reads with at least one base overlapping the gene from the start to the end coordinates are retrieved. CIGAR strings are then used to find the alignments blocks. Junction reads are identified by the presence of at least one'N' letter in the CIGAR. Junction reads were filtered if:

- More than 10% of soft-clipping was detected in the alignment (it should not be the case with TopHat).
- An indel was close to the junction site, as it would make the junction position uncertain.

Junction read alignments are then processed block by block sequentially from left to right. Alignment blocks under 4 bp on read extremities are removed from the reads as we considered it is not sufficient to identify correctly the mapping localization. Then each block is compared to FasterDB annotations to check if the block boundaries correspond to known exons annotated in FasterDB, or to a putative new acceptor or donor site. First and last alignment blocks for each read must overlap one and only one exon for a read to be considered. For the inner blocks, if alignment blocks map to a succession of exons and introns, it is considered as an intron retention. For the acceptors and donors, we also added a supplementary filter. If a new donor is identified within a junction, we check if the junction also has an acceptor identified of the same length +/−1bp on the other side of the junction, showing most probably a problem of mapping. Once all the blocks are identified, the block annotations are used to annotate putative alternative splicing events: alternative skipped exon, multiple exon skipping, acceptor, or donor sites.

Once all the junction reads are processed, the alternative splicing events identified are pooled and the reads participating to each event are quantified, as well as the known exon-exon junction. If an exon-exon junction

is annotated with multiple known acceptors and/or donors, all the possible junction reads are quantified and summed up. To fasten the quantification step, a junction coordinate file with the corresponding read numbers is produced from the read alignment using the same filters than described above and will be used for all the quantification tools: junction, exon skipping, acceptor and donor.

A challenge in defining the alternative skipped exon events is to identify the flanking exons. In the first version of FaRLine, these flankings exons were defined as the closest annotated genomic exons. This rule led to miss a lot of ASE events. Therefore, to define the flanking exons, we now use the information contained in the transcripts and in the reads. We consider each junction which skips an exon and is covered by at least one read. If this junction is annotated in the transcripts, we extract all annotated events containing this junction. Else, we annotate the event with the longest covered inclusion isoform. It allows FaRLine to be more robust to the incompleteness of the annotation compared to other methods, like MISO (Supplementary Figure S6). Panel C of Supplementary Figure S8 gives an example of an ASE reported by FaRLine but not by MISO because the exclusion isoform is not annotated in the transcripts.

Comparison with STAR: We also mapped the reads with STAR, ran FaRLine on these alignments and compared the predicted skipped exons with KisSplice. The main results are similar to what we found with TopHat. Indeed, without any filter, 69% of ASE annotated by KisSplice are also found by FaRLine and 24% of FaRLine's event by KisSplice (compared to 68% and 24% respectively for the mapping with TopHat). When we filter out the events with an unfrequent variant, we show that approximately 70% of predicted ASE are found by both approaches.

Quantification and differential analysis. Both pipelines perform ASE detection and quantification. The quantification step was done similarly in the two pipelines where only the junction reads were taken into account. To evaluate if using exonic reads in the quantification could increase the accuracy of our methods, we ran KisSplice on the MCF-7 dataset with the option –exonic reads set to on. In doing so, only the inclusion rate of the AS events changes. When comparing usage of only junction reads to usage of both junction and exonic reads, we observed that the p-values calculated strongly correlate as shown in Supplementary Figure S16. We found that some AS events became significant upon the addition of exonic reads but the opposite also happened. Inspection of these events revealed that many are borderline cases, where the p-value is close, but slightly above 5%. A manual inspection of the AS events with a very different p-value upon addition of exonic reads revealed that they correspond to exons overlapping alternative first or last exons (see *STARD4*, Supplementary Figure S17A) or novel exons located in poorly spliced introns (see *PANK2* and *PRRC2B*, Supplementary Figure S17 B and C). Overall, we concluded that exonic reads can bring some statistical power in cases where the skipped exon does not overlap with any other event. In case of more complex events, exonic reads tend to "pollute" the pairwise comparison.

The last step of the pipelines is the differential analysis of the expression levels of the variants. This task is performed using the kissDE[24] R package, which takes as input a table of read counts as in Supplementary Figure S18, and outputs a p-value and a DeltaPSI (Percent Spliced In).

Our statistical analysis adopted the framework of count regression with Negative Binomial distribution. We considered a 2-way design with interaction, with *isoforms* and *experimental conditions* as main effects. Following the Generalized Linear Model framework, the expected intensity of the signal was denoted by λ_{ijk} and was decomposed as:

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad (1)$$

where μ is the local mean expression of the gene, α_i the contribution of splicing variant i on the expression, β_j the contribution of condition j to the total expression, and $(\alpha\beta)_{ij}$ the interaction term. The target hypothesis was H_0: $\{(\alpha\beta)_{ij} = 0\}$ *i.e.* no interaction between the variant and the condition. If this interaction term is not null, a differential usage of a variant across conditions occurred. The test was performed using a Likelihood Ratio Test with one degree of freedom. To account for multiple testing, p-values were adjusted with a 5% false discovery rate (FDR) following a Benjamini-Hochberg procedure[34].

In addition to adjusted p-values, we report a measure of the magnitude of the effect. The measure we provide is based on the Percent Spliced In (PSI):

$$PSI_{condition} = \frac{counts_{variant1}}{counts_{variant1} + counts_{variant2}} \quad (2)$$

If counts for a variant are below a threshold, then the PSI is not calculated. This prevents from over-interpreting large magnitudes derived from low counts. When several replicates are available for a condition, then a PSI is computed for each replicate, and we calculate their mean.

Finally, we output the DeltaPSI:

$$DeltaPSI = PSI_{condition1} - PSI_{condition2} \quad (3)$$

unless one of the mean PSI of a condition could not be estimated. The higher the DeltaPSI, the stronger the effect. In practice, we consider only DeltaPSI larger than 0.1, a threshold below which it is difficult to perform any experimental validation.

SK-N-SH dataset. We downloaded a total of 959 M reads from http://genome.crg.es/encode_RNA_dashboard/hg19/35. They correspond to long polyA+ RNAs generated by the Gingeras lab, and are also accessible with the following accession numbers (ENCSR000CPN - SRA: SRR315315, SRR315316 and ENCSR000CTT

-SRA: SRR534309, SRR534310). For cell lines treated by retinoic acid, the reads were 76nt long, while they were 100nt long for the non treated cells. Hence we trimmed all reads to 76nt.

MCF-7 dataset. MCF-7 were transfected (two biological replicates) with siRNA targeting both DDX5 and DDX17 RNA helicases, and total RNA were extracted as described previously[36]. cDNA synthesis was made using the TruSeq Stranded Total RNA protocol after Ribo-Zero Gold-mediated elimination of ribosomal RNA (Beckman Coulter Genomics). High throughput sequencing (2×125 bp) was carried out on an Illumina HiSeq 2500 platform (Beckman Coulter Genomics), generating between 45 and 50 millions of paired-end pairs of reads. Raw datasets are available on GEO under the accession number GSE94372.

Reads were trimmed according to standard quality control filters using prinseq[37] and adapter were removed using cutadapt[38]. The resulting reads had length between 25 and 125nt. Because MISO is unable to deal with reads of unequal length, we selected only reads with length larger than 100nt (87% of the reads) and trimmed longer reads to 100nt.

Computational requirements, software availability and reproducibility of the results. FaRLine took 45 hours and 10 Go of RAM. The time-limiting step was TopHat2, which took 41 hours, even parallelised on 8 cores. When STAR was tested instead of TopHat2, it took 4 hours, but 30 Go of RAM. KisSplice took 30 hours and 10 Go of RAM. The RAM-limiting step was STAR which took 30Go of RAM. All the steps of the pipelines can be reproduced using the following tutorial:

http://kissplice.prabi.fr/pipeline_ks_farline.

Experimental Validation. SK-N-SH cells were purchased from the American Type Culture Collection (ATCC) and cultured using EMEM medium (ATCC) complemented with 10% FBS (Thermo Fisher Scientific). Cells were differentiated for 48 h using 6 μM of all-trans retinoic acid (Sigma-Aldrich).

After harvesting, total RNA were extracted using Tripure isolation reagent (Sigma-Aldrich), treated with DNase I (DNAfree, Ambion) for 30 min at 37 °C and reverse-transcribed (RT) using M-MLV reverse transcriptase and random primers (Invitrogen). Before PCR, all RT reaction mixtures were diluted at 2.5 ng μL of initial RNA. PCR reactions were performed using GoTaq polymerase (Promega).

MCF7 cells were cultured as described in[36]. RT-PCRs were performed using the same protocol as for SK-N-SH cells.

References

1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–1415 (2008).
2. Wang, E. T. *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
3. Scotti, M. M. & Swanson, M. S. Rna mis-splicing in disease. *Nature Reviews Genetics* **17**, 19–32 (2016).
4. Edery, P. *et al*. Association of tals developmental disorder with defect in minor splicing component u4atac snrna. *Science* **332**, 240–243 (2011).
5. David, C. J. & Manley, J. L. Alternative pre-mrna splicing regulation in cancer: pathways and programs unhinged. *Genes & development* **24**, 2343–2364 (2010).
6. Trapnell, C. *et al*. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols* **7**, 562–578 (2012).
7. Wang, K. *et al*. Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic acids research* **38**, e178–e178 (2010).
8. Robertson, G. *et al*. *De novo* assembly and analysis of rna-seq data. *Nature methods* **7**, 909–912 (2010).
9. Steijger, T. *et al*. Assessment of transcript reconstruction methods for rna-seq. *Nature methods* **10**, 1177–1184 (2013).
10. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome research* **22**, 2008–17 (2012).
11. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods* **7**, 1009–1015 (2010).
12. Shen, S. *et al*. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Research* e61–e61 (2012).
13. Sacomoto, G. A. T. *et al*. KISSPLICE: *de-novo* calling alternative splicing events from RNA-seq data. *BMC bioinformatics* **13**(Suppl 6), S5 (2012).
14. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nature Reviews Genetics* **12**, 671–682 (2011).
15. Dargahi, D. *et al*. A pan-cancer analysis of alternative splicing events reveals novel tumor-associated splice variants of matriptase. *Cancer informatics* **13**, 167 (2014).
16. Freyermuth, F. *et al*. Splicing misregulation of scn5a contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy. *Nature communications* **7** (2016).
17. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology* **29**, 644 (2011).
18. Kopelman, N. M., Lancet, D. & Yanai, I. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37**, 588–589 (2005).
19. Roux, J. & Robinson-Rechavi, M. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome research* **21**, 357–363 (2011).
20. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nature Reviews Genetics* **3**, 370–379 (2002).
21. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3'splice-site selection in alu exons. *Science* **300**, 1288–1291 (2003).
22. Sorek, R. *et al*. Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons. *Molecular cell* **14**, 221–231 (2004).
23. Franz, M. *et al*. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* **32**, 309–311, https://doi.org/10.1093/bioinformatics/btv557/oup/backfile/content_public/journal/bioinformatics/32/2/10.1093_bioinformatics_btv557/3/btv557.pdf (2016).
24. Lopez-Maestre, H. *et al*. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research* **44**, e148–e148 (2016).

25. Poursani, E. M., Soltani, B. M. & Mowla, S. J. Differential expression of oct4 pseudogenes in pluripotent and tumor cell lines. *Cell Journal (Yakhteh)* **18**, 28 (2016).
26. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* **17**, 1 (2016).
27. Shen, S. *et al*. Widespread establishment and regulatory impact of alu exons in human genes. *Proceedings of the National Academy of Sciences* **108**, 2837–2842 (2011).
28. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 9869–74 (2014).
29. Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome biology* **16**, 204 (2015).
30. Mallinjoud, P. *et al*. Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome research* **24**, 511–521 (2014).
31. Laustriat, D. *et al*. *In Vitro* and *In Vivo* Modulation of Alternative Splicing by the Biguanide Metformin. *Molecular Therapy. Nucleic Acids* **4**, e262 (2015).
32. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust *de novo* rna-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092 (2012).
33. Li, H. *et al*. The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).
34. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 289–300 (1995).
35. Djebali, S. *et al*. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
36. Dardenne, E. *et al*. RNA Helicases DDX5 and DDX17 Dynamically Orchestrate Transcription, miRNA, and Splicing Programs in Cell Differentiation. *Cell Reports* (2014).
37. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)* PMID: 21278185. **27**, 863–864 (2011).
38. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17** (2011).

Acknowledgements

This work was performed on the computing facilities of the computing center LBBE/PRABI and the PSMN (Pole Scientifique de Modelisation Numerique) computing center of ENS de Lyon. This work was funded by the ANR-12-BS02-0008 (Colib'read) by the ABS4NGS ANR project (ANR-11-BINF-0001-06), Action n3.6 Plan Cancer 2009–2013, Fondation ARC (Programme Labellisé Fondation ARC 2014, PGA120140200853) and INCa (2014-154). Doctoral fellowships from ARC 1 - Région Rhône-Alpes (C.B.P), Science Without Borders - CNPq - Brazil (L.L. - grant process number 203362/2014-4), ARS RhÃ´ne-Alpes (A.R.) and post-doctoral fellowships from Fondation ARC (M.P.L).

Author Contributions

V.L. and D.A. designed the study. C.B.P., E.C. and J.B.C. developed FaRLine. L.L. and G.S. significantly improved the scalability of KisSplice, C.M., A.C. and V.L. developed KisSplice2RefGenome. C.B.P., L.L. and V.L. compared the two pipelines and classified the instance types. L.L. developed the supporting webpage. C.B.P. and C.F.B. planned the experimental validations. M.P.L., A.R., S.T., L.D. performed the experimental validations. C.B.P., D.A. and V.L. wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-21770-7.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data

Clara Benoit-Pilven[1], Camille Marchet[3], Emilie Chautard[1,2], Leandro Lima[2], Marie-Pierre Lambert[1], Gustavo Sacomoto[2], Amandine Rey[1], Audric Cologne[2], Sophie Terrone[1], Louis Dulaurier[1], Jean-Baptiste Claude[1], Cyril F. Bourgeois[1], Didier Auboeuf[1,*], and Vincent Lacroix[2,*]

[1]Université de Lyon, ENS de Lyon, Université Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France

[2]Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France. EPI ERABLE - Inria Grenoble - Rhône-Alpes

[3]IRISA Inria Rennes Bretagne Atlantique CNRS UMR 6074, Université Rennes 1, GenScale team, Rennes, 263 Avenue Général Leclerc, Rennes, France

[*]Corresponding authors : Vincent Lacroix (Vincent.lacroix@univ-lyon1.fr) and Didier Auboeuf (Didier.auboeuf@inserm.fr)

Supplementary methods :

KisSplice

Alternative splicing events are bubbles in the DBG

Supplementary figure S13 gives a schematic example of two alternative transcripts which differ by the inclusion of one exon. For the sake of simplicity, the example is given for words of length 3, but the reasoning holds for any word length. Each distinct word of length k is called a k-mer and corresponds to a node of the DBG. There is a directed edge from a node u to a node v if the last $k-1$ nucleotides of u are identical to the first $k-1$ nucleotides of v. Each transcript will therefore correspond to a path in the DBG. A pair of internally node-disjoint paths with a common source and target is called a bubble. The smaller path of the bubble corresponds to the exclusion isoform and is composed of all k-mers which overlap the junction between the exons flanking the skipped exon. It is therefore usually composed of $k-1$ k-mers. In the special case where the skipped exon shares a prefix with its 3' flanking exon, or a suffix with its 5' flanking exon, then the lower path is composed of less than $k-1$ k-mers and the k-mer which is the source (resp. target) does not correspond anymore to an exonic k-mer, but to a junction k-mer.

In practice, the DBG is built from the reads, not from the transcripts. The reads stem from possibly all genes expressed in the studied conditions.

Two difficulties arise: reads contain sequencing errors, and repeats may be shared across genes.

Dealing with sequencing errors

As originally described in[1] and later in[2], sequencing errors generate recognisable structures in De Bruijn graphs, which can be identified and removed. Their systematic removal however prevents assemblers from studying SNPs. A compromise consists in discarding rare k-mers from the graph. This is the strategy we use in KisSplice, where we remove all k-mers seen only once. This idea is however not sufficient in the context of transcriptome assembly, where the coverage is very uneven and mostly reflects expression levels. For highly expressed genes, several reads may have errors at the same site, generating k-mers with a coverage larger than an absolute threshold. We therefore also use a relative cut-off, which we set to 2%. These cut-offs we introduce to remove sequencing errors have an impact on the running time and on the sensitivity. Decreasing them allows to discover rarer isoforms, at the expense of a longer running time.

Dealing with repeats

Repeats are notoriously difficult to assemble in DNAseq data, and were initially thought to be much less problematic in RNAseq, since they are mostly located in introns and intergenic regions. In practice, mRNA extraction protocols are not perfect, and a fraction of pre-mRNA remains (typically 5% for total polyA+ RNA[3]). Each intron is covered by few reads, but if a repeat is present in many introns, then this repeat will obtain a high coverage. If, in addition, the multiple copies of the repeat are not identical, the repeat family will correspond to a very dense subgraph in the De Bruijn graph built from the reads. The traversal of such subgraph to enumerate all the bubbles it contains is long and mostly fruitless, although some true AS events flanked by

repeats may be trapped in these subgraphs. We showed in[4] that an effective strategy to deal with this issue is to enumerate only bubbles which have at most *b* branches. In practice, we set *b* to 5. Increasing *b* will increase the running time, but allow to find more repeat-associated alternative splicing events. Bubbles which do not correspond to true AS events can be filtered out at the mapping step.

MISO

MISO[5] was run in "exon-centric" mode with default parameter. We first generated from the EnsEMBL r75 gff file the alternative event annotation file requested by MISO using `rnaseqlib`. The mapping step was done exactly the same as for FARLINE with Tophat-2.0.11[6], except that the replicates of each condition were merge together because MISO does not accept biological replicates. We then run all MISO scripts with default parameters. Finally, we filtered the differentially changing events with the `filter_events` script using the following parameters :
`--num-sum-inc-exc 10 --delta-psi 0.1 --bayes-factor 20`.

Cufflinks

Cufflinks[6] was run on the same alignment files used in FARLINE using annotation as a guide with the following parameters :
`-g <EnsEMBL r75 gff file> -b <hg19 genome> -u -p 16`.
When an annotation is given as a guide to Cufflinks, some faux-reads are introduced to support all transcripts present in the annotation. Because it can annotate transcripts even if there are not expressed in the samples, for the rest of the analysis, we decide to consider only the reconstructed transcripts supported by real reads.
Then, the AS events were retrieved from the reconstructed transcripts using the FARLINE annotation script.

Trinity

Trinity[7] was run with the following parameters :
`--max_memory 110G --CPU 16 --min_kmer_cov 2 --seqType fq --SS_lib_type RF`. In order to retrieve the bubbles from Trinity's output file, we parsed the transcripts' headers by firstly partitioning the reconstructed transcripts into disjoint sets, where each set is a predicted gene. Then, for each such set, the bubbles were found by processing the nodes' identifiers used to build each isoform.

References

1. Pevzner, P. A., Tang, H. & Tesler, G. De novo repeat classification and fragment assembly. *Genome research* **14**, 1786–1796 (2004).

2. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research* **18**, 821–829 (2008).

3. Tilgner, H. *et al.* Deep sequencing of subcellular rna fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas. *Genome research* **22**, 1616–1625 (2012).

4. Sacomoto, G. *et al.* Navigating in a Sea of Repeats in RNA-seq without Drowning. *Lect.* **8701**, 82–96 (2014).

5. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat. methods* **7**, 1009–1015 (2010).

6. Trapnell, C. *et al.* Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat. protocols* **7**, 562–578 (2012).

7. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nat. biotechnology* **29**, 644 (2011).

Supplementary figures :

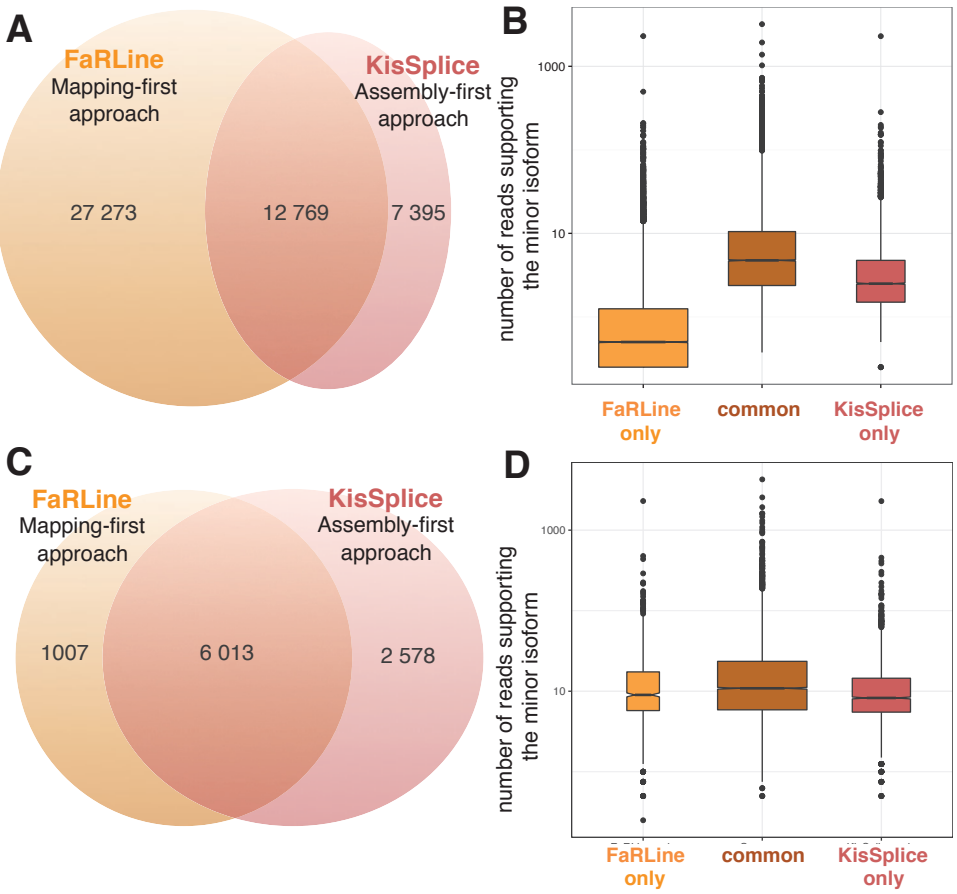

Figure S1. Comparison of the ASE identified by the assembly-first and mapping-first pipelines on MCF-7 dataset. A) Venn diagram of ASEs identified by the two pipelines. FARLINE detected many more events than KISSPLICE. 63% of ASE found by KISSPLICE were also found by FARLINE and 32% of ASE detected by FARLINE were also found by KISSPLICE. B) Boxplot of the expression of the minor isoform in the 3 categories defined in the Venn diagram of panel A: ASE identified only by FARLINE, ASE identified by both pipelines and ASE identified only by KISSPLICE. The number of reads supporting the minor isoform of the ASE identified by FARLINE is globally much lower. Many isoforms are supported by less than 5 reads. C) Venn diagram of ASEs found by the two pipelines after filtering out the poorly expressed isoforms. The common events represent a larger proportion than before filtering: 86% of the ASE annotated by FARLINE and 70% of the ASE annotated by KISSPLICE. D) Boxplot of the expression of the minor isoform in the 3 categories defined in the Venn diagram of panel C: ASE identified only by FARLINE, ASE identified by both pipelines and ASE identified only by KISSPLICE. The distribution of the number of reads supporting the minor isoform is similar for the 3 categories with highly expressed variants in each category.

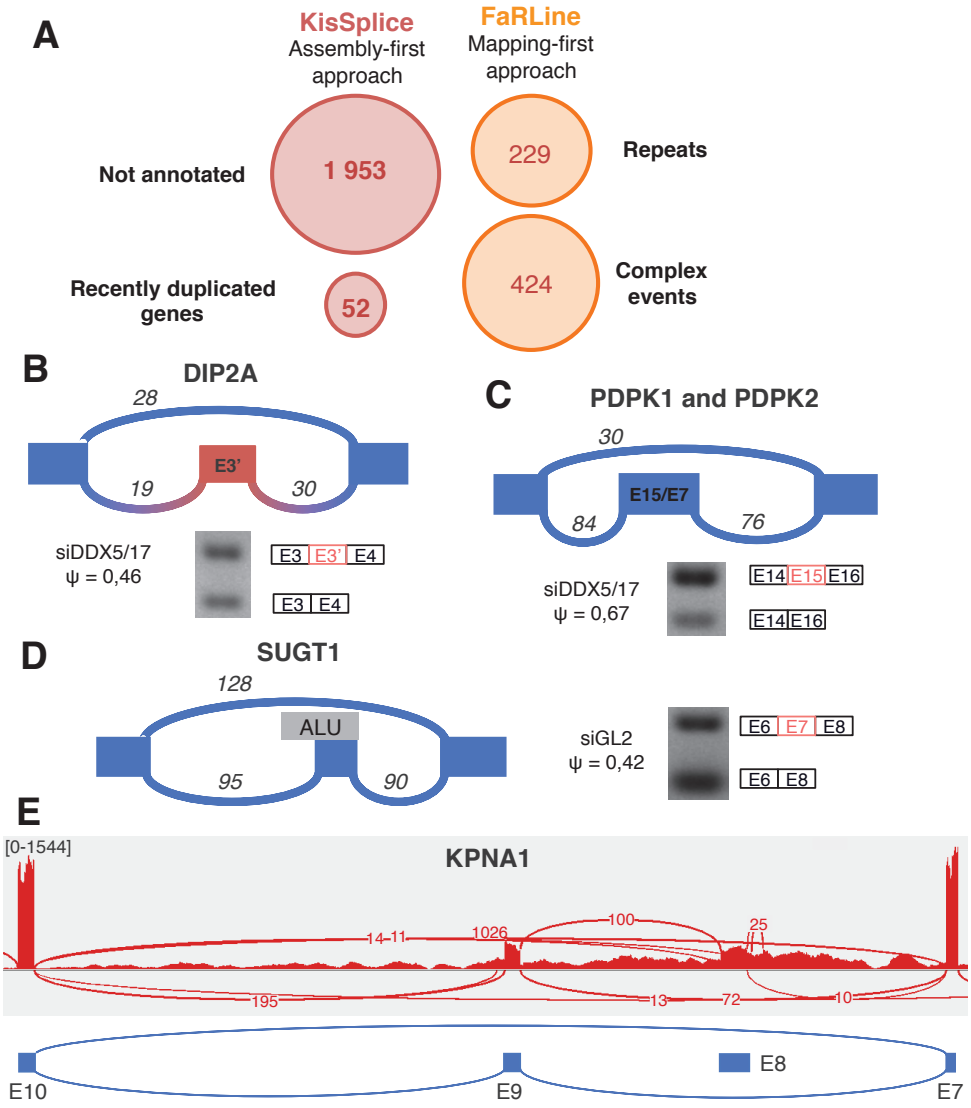

Figure S2. A) Main categories identified explaining why some exons are detected by only one method. Numbers for MCF-7 dataset. B) The exon in intron 3 of the *DIP2A* gene is an example of an exon not annotated in EnsEMBL r75. This event was identified by KISSPLICE but not by FARLINE. C) *PDPK1* and *PDPK2* are 2 paralog genes. KISSPLICE detected 2 isoforms that could be produced by these 2 genes. FARLINE did not detect any event in either of these genes. The exon skipped is exon 15 in *PDPK1* (corresponding to exon 7 in *PDPK2*). C) Exon 7 of the *SUGT1* gene is an example of exon skipping overlapping an Alu element identified only by FARLINE. The events in panel A to C were validated by RT-PCR. E) The *KPNA1* gene contains a complex event with more than 5 branches inside the bubble. This event was detected by FARLINEbut not by KISSPLICE

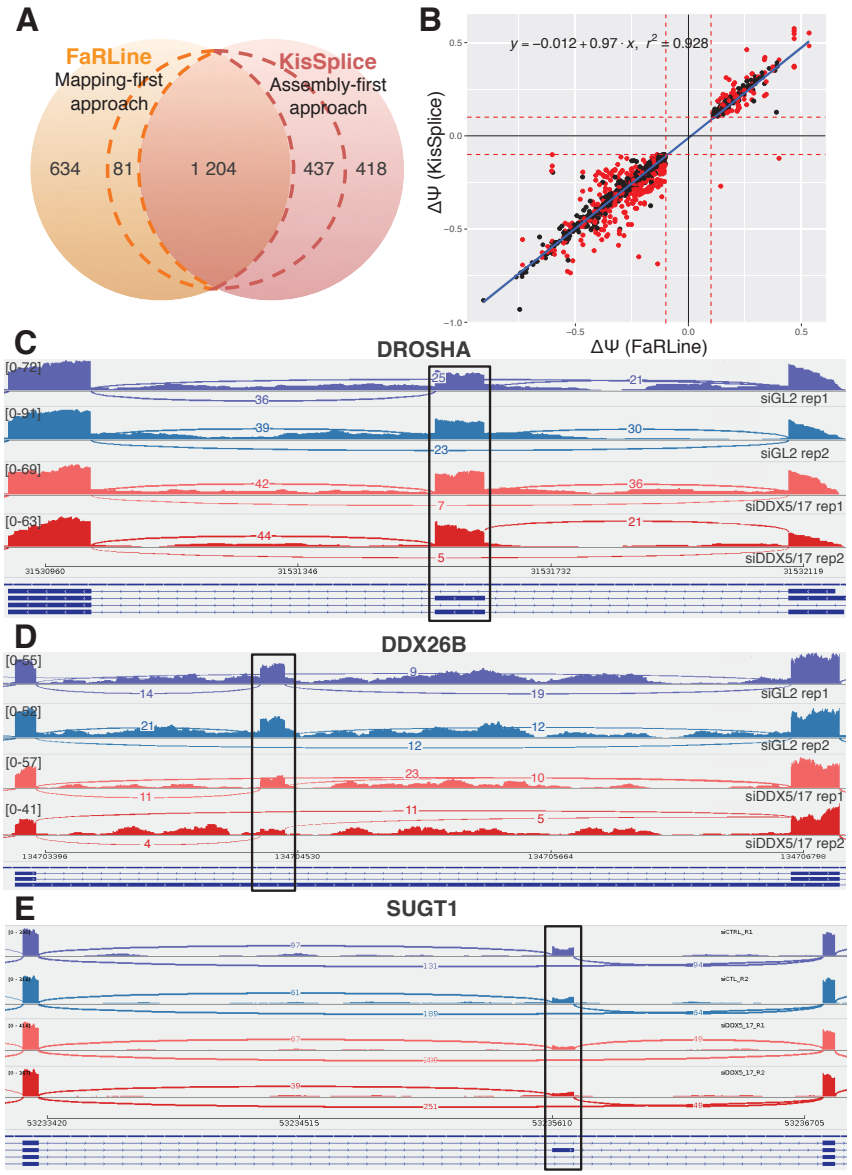

Figure S3. A) Condition-specific variants found by FARLINE, KISSPLICE or both methods in MCF-7 dataset. Within dashed lines are events identified by both approaches but detected as condition-specific by only one approach. B) DeltaPSI as estimated by KISSPLICE and FARLINE, for events identified by both methods. The red dots represent complex events for which KISSPLICE found at least 2 'bubbles'. C) Exon 2 of *DROSHA* is an example of regulated ASE found by both approaches. D) A new exon in intron 10 of the *DDX26B* gene is found only by KISSPLICE. The inclusion rate of this exon is differentially regulated between the 2 experimental conditions. E) Because exon 7 of the *SUGT1* gene is an exonised Alu element, only FARLINE identified this event. Moreover this exon is significantly more included in the control cells (expressing DDX5 and DDX17) when compared to the DDX5/DDX17 depleted cells.

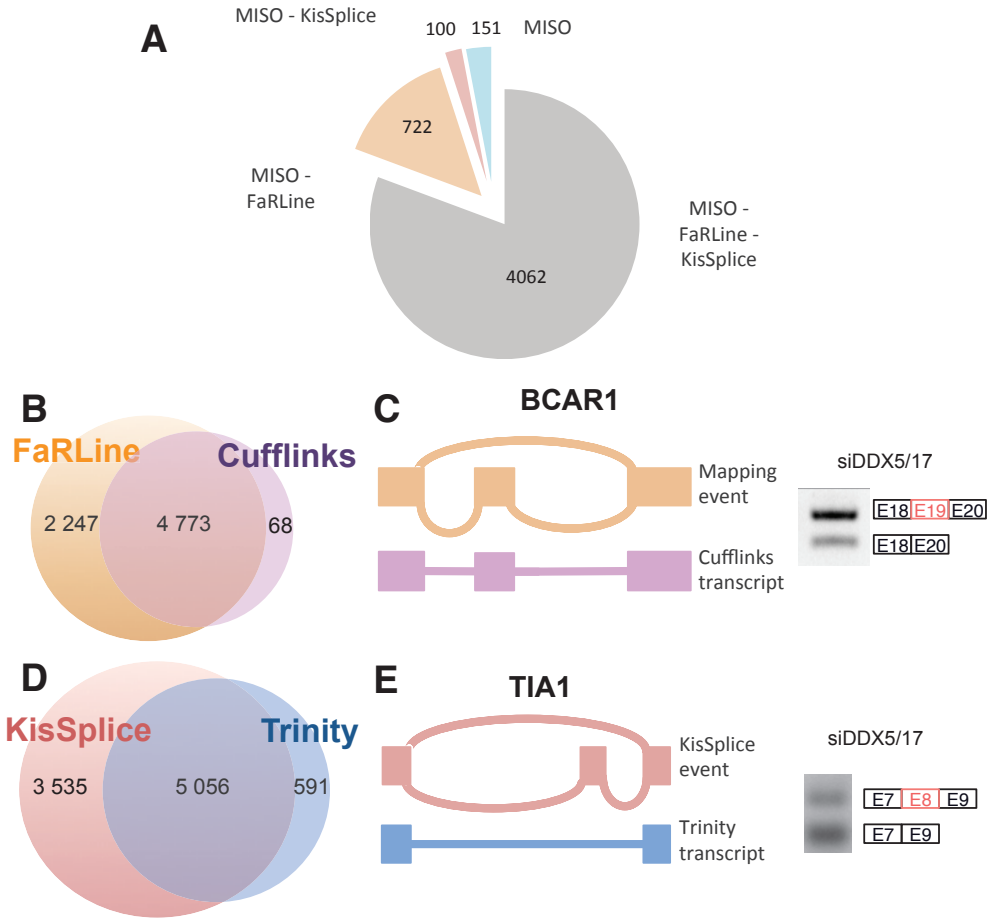

Figure S4. A) 81% of ASE found by MISO are also annotated by FARLINE and KISSPLICE. 14% of MISO's ASE are also annotated by FARLINE while only 2% of MISO's ASE are also annotated by KISSPLICE. Finally, only 3% of these ASEs are only annotated by MISO. B) Most of the events annotated by Cufflinks are found by FARLINE. C) *BCAR1* exon 19 is an example of an ASE annotated by FARLINE but not by Cufflinks. Indeed, only the inclusion isoform was identified by Cufflinks. D) Most of the events annotated by Trinity are also found by KISSPLICE. However half of the ASE annotated by KISSPLICE are not found by the global assembler Trinity. E) KISSPLICE annotates an ASE in the *TIA1* gene, while Trinity only identified the exclusion variant. The events in panels C and E have been validated by RT-PCR.

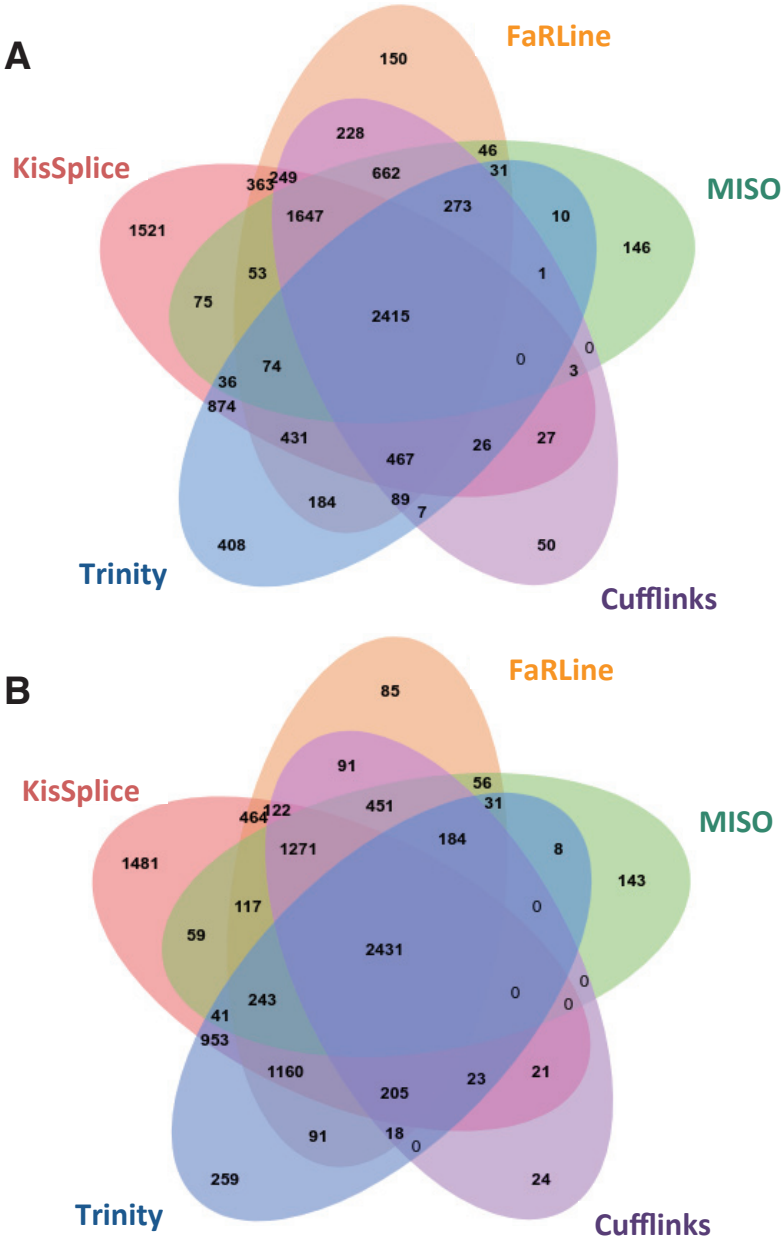

Figure S5. Venn diagram of the comparison of the five methods : KisSplice, FaRLine, MISO, Cufflinks and Trinity, on the SK-N-SH dataset (A) and on MCF-7 dataset (B). The total number of annotated splicing events predicted by at least one method, with the minor isoform being supported by at least 5 reads is 10546. The largest overlaps are 2415 (all methods), 1647 (all methods but Trinity), 874 (KisSplice-Trinity), 662 (FarLine-MISO-Cufflinks). As expected, Trinity is the least sensitive method. We also observe that the three mapping-first approaches (FarLine, MISO and Cufflinks) have a very large number of common candidates, 662 of which are not found by the two assembly-first approaches (KisSplice and Trinity). Conversely, the two assembly-first approaches have a very large number of common candidates, 874 of which are not found by the three mapping-first approaches. Similar numbers are found for the MCF-7 dataset. These results support the main conclusion of this paper.

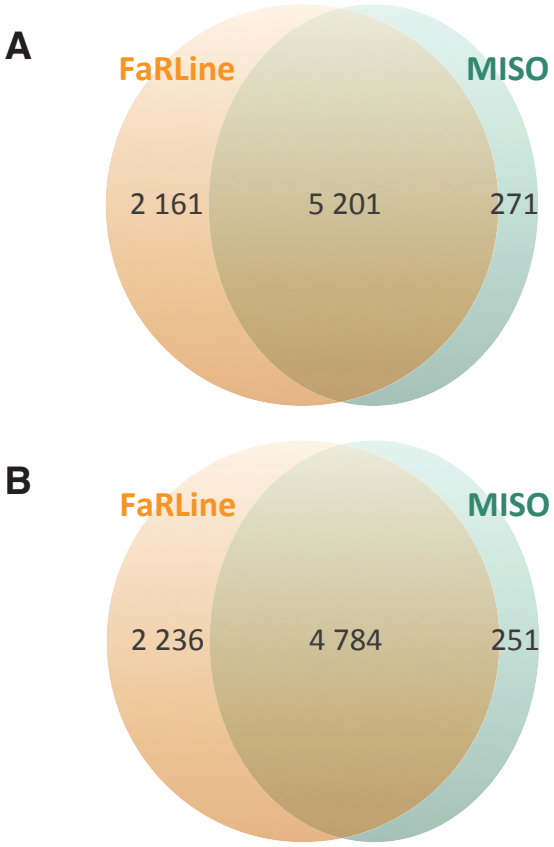

Figure S6. Venn diagram of the comparison of FaRLine and MISO on SK-N-SH dataset (A) and on MCF-7 dataset (B).

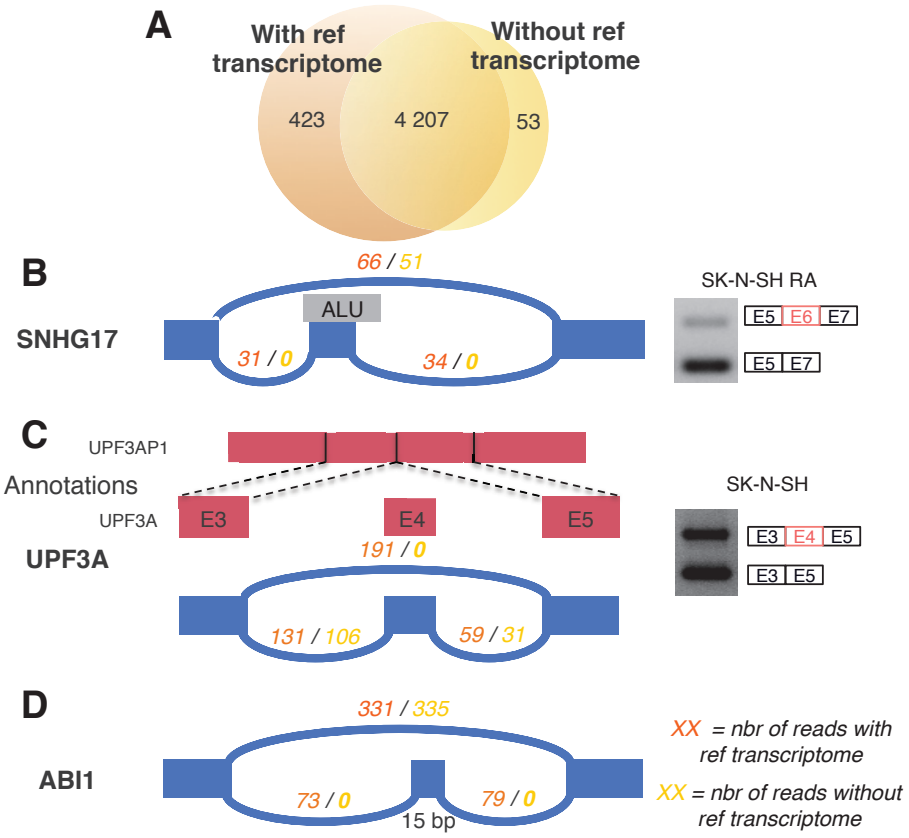

Figure S7. Comparison of the mapping-first approach FARLINE with or without an annotation provided to the mapper (i.e. with/without reference transcriptome) on the SK-N-SH dataset. A) More ASE are annotated when an annotation available. Panels B to D show examples of events only found by the mapping-first method when an annotation is provided to the mapper. B) The first category, represented by the *SNHG17* gene, includes exons containing repeats like ALU elements. C) Genes with a retrotransposed pseudogene, as *UPF3A*, represent the second category and are more difficult to find when no annotation is available. D) Short exons (less than 20 bp), like exon 5 of the *ABI1* gene, compose the third category.

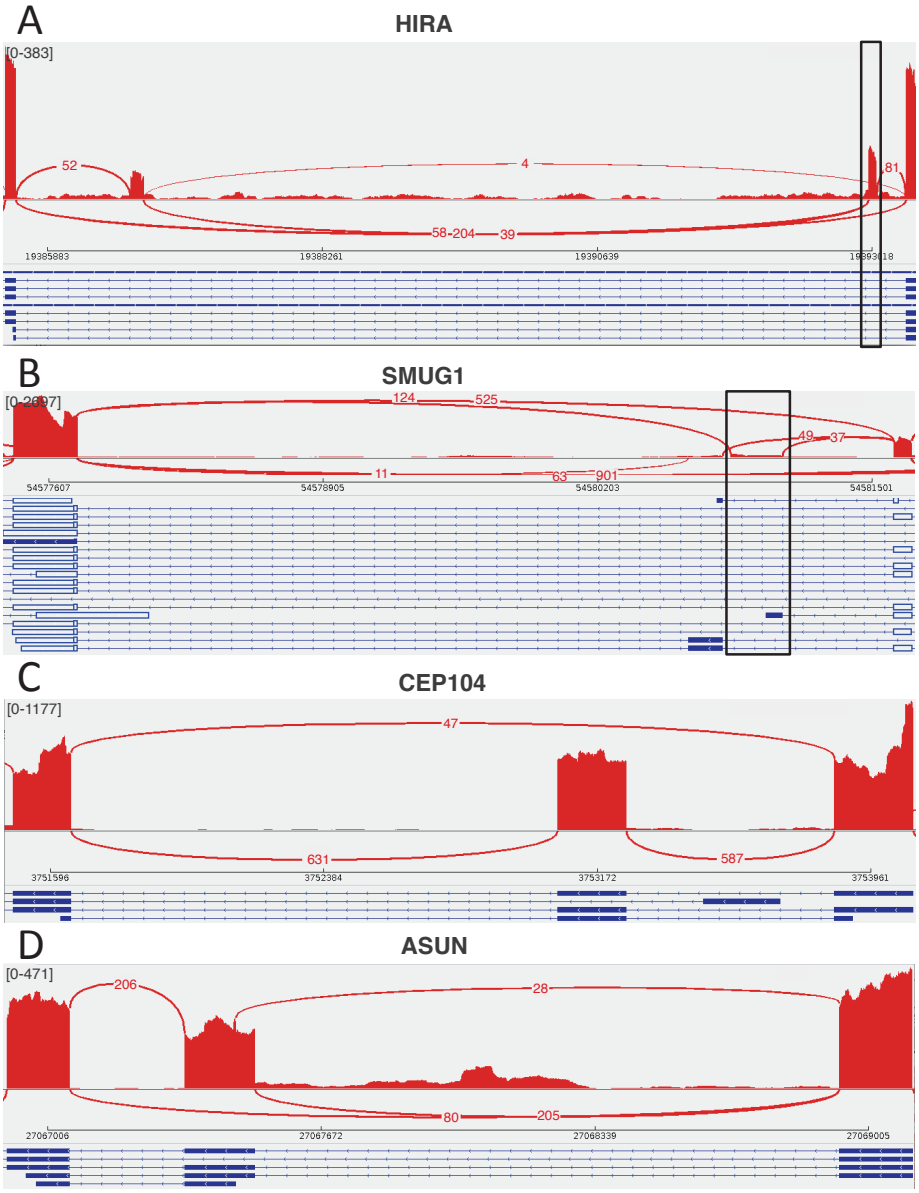

Figure S8. Examples of exon skipping inside a complex event. A) A new exon in intron 8 (black box) of *HIRA* gene is reported as skipped by KISSPLICE with exons 8 and 9 as flanking exons. B) The exon 5 of *SMUG1* gene is reported as skipped by KISSPLICE with exons 4 and 7 as flanking exons. This event is not found by FARLINE because the inclusion isoform is not annotated in the transcrits. C) Exon 12 of gene *CEP104* is reported as skipped by FARLINE even if the exclusion isoform is not present in the annotation. However, MISO does not find this exon skipping. D) Example of an exon skipping with two alternative donor sites in *ASUN* gene. It is reported as one event by FARLINE and two events by KISSPLICE.

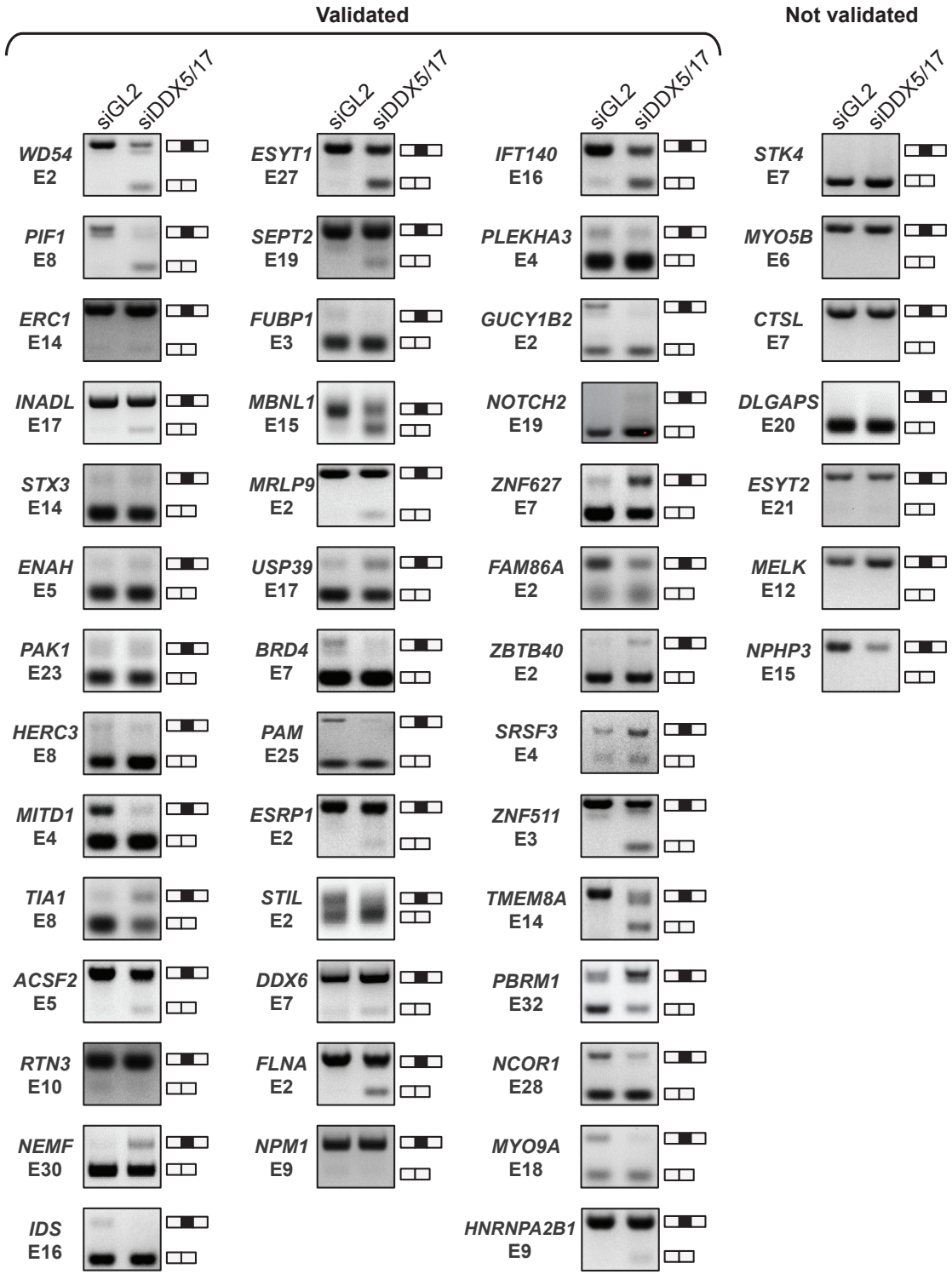

Figure S9. RT-PCR validations of events found by both approaches in the MCF-7 dataset. 41 out of the 48 events were validated (both the inclusion and the exclusion variant were amplified by RT-PCR). In some cases, there were additional PCR products (marked as '*') suggesting the existence of additional variants.

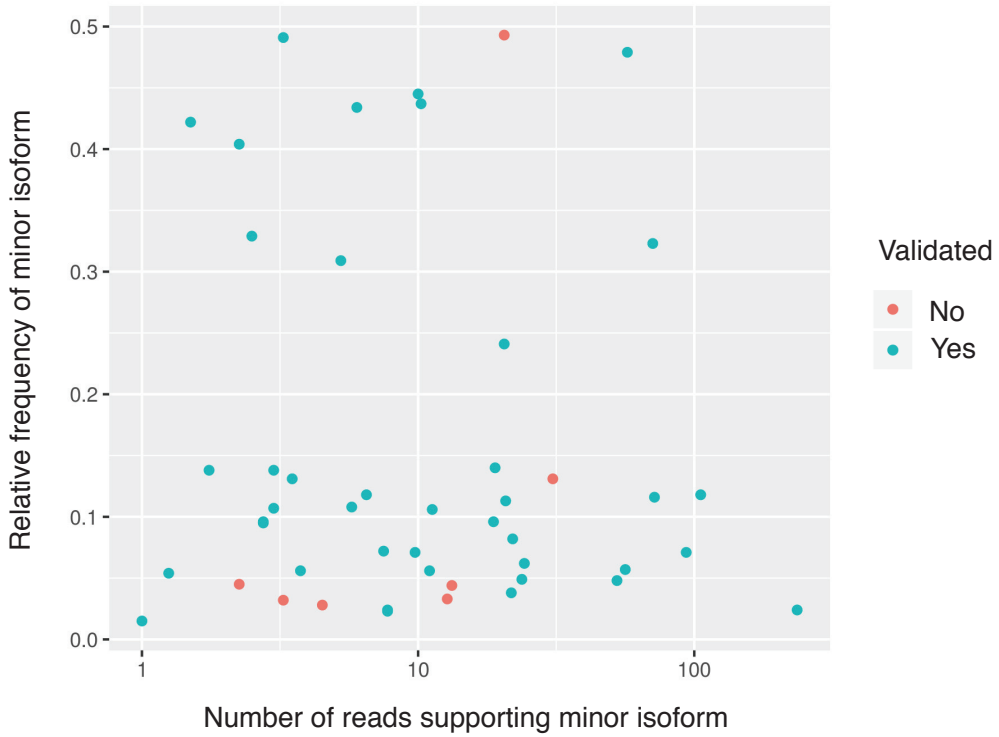

Figure S10. Repartition of validated and non validated ASEs according to number of reads supporting the minor isoform, and relative frequency of the minor isoform (i.e. number or reads of the minor isoform / number of reads supporting both isoforms). The X axis is in log scale. Most of the non validated cases have relative frequency of their minor isoform lower than 10%.

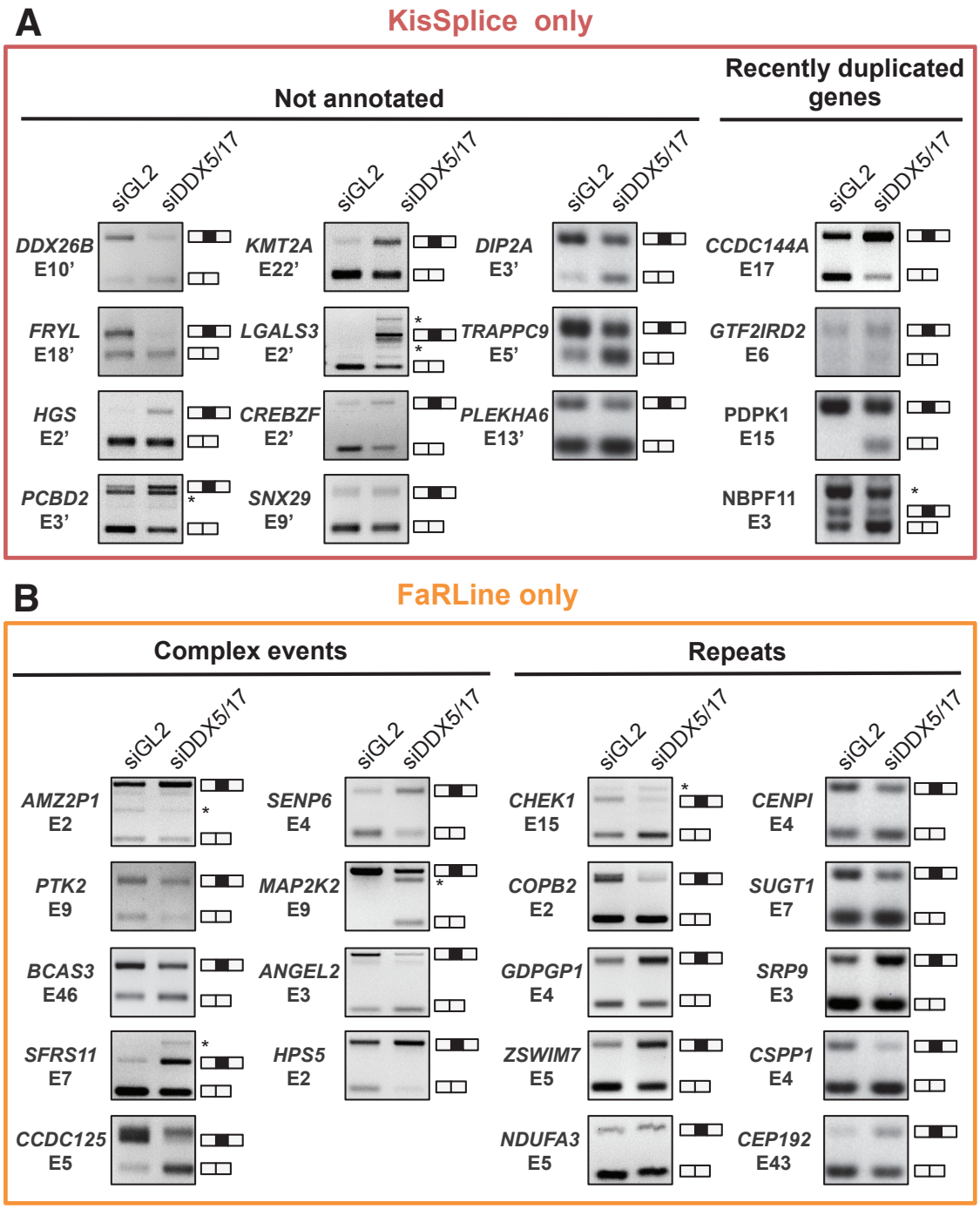

Figure S11. RT-PCR validations of events found only by KISSPLICE (A) and only by FARLINE (B) in the MCF-7 dataset. These ASEs were selected from the 4 main categories shown in Figure 3 and Supplementary Figure S2. All of them were validated.

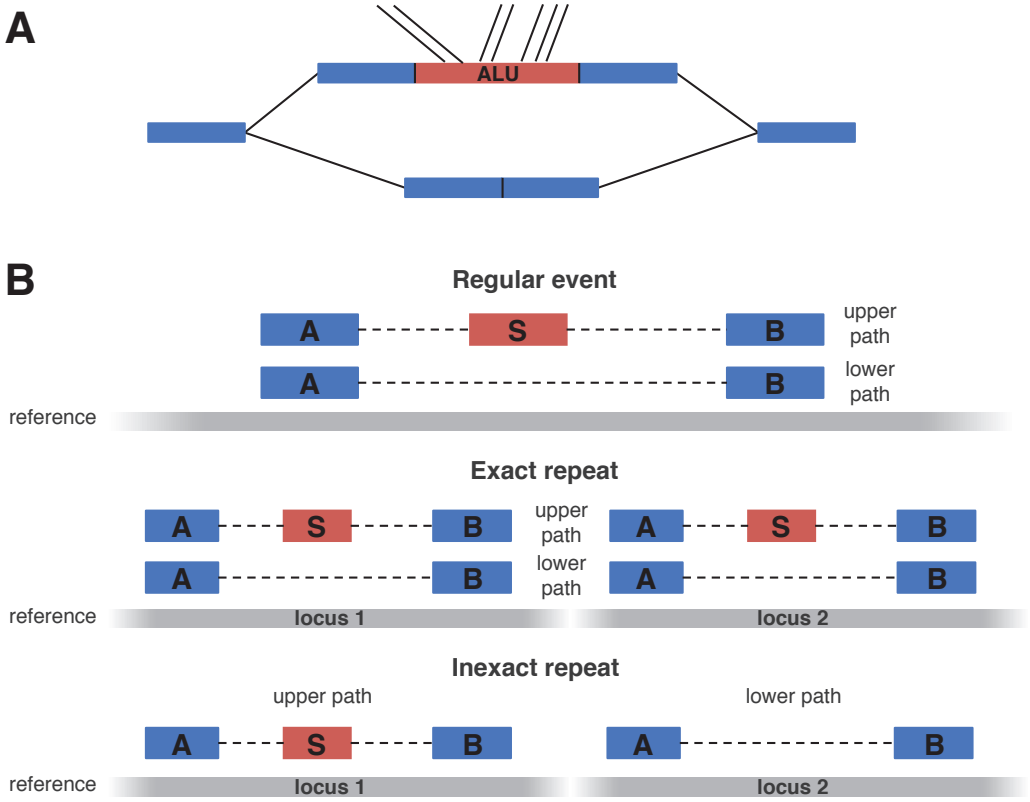

Figure S12. Dealing with repeats in KISSPLICEand KISSPLICE2REFGENOME. A) Example of a bubble containing an Alu. Repeated events such as Alu are expected to be present in several copies in the reads. Thus, when the graph is constructed, edges link different copies of Alu. Because a bubble with more than 5 edges within one of its paths is not enumerated by KISSPLICE, this case is not annotated by the assembly-first approach. B) In KISSPLICE2REFGENOME, if the two variants (i.e. paths) both map on different copies (exact repeat), we classify it as a recent paralog. On the contrary if each variant maps on a different locus, we consider the event as coming from an inexact repeat. This category represents mostly paralogs that have diverged.

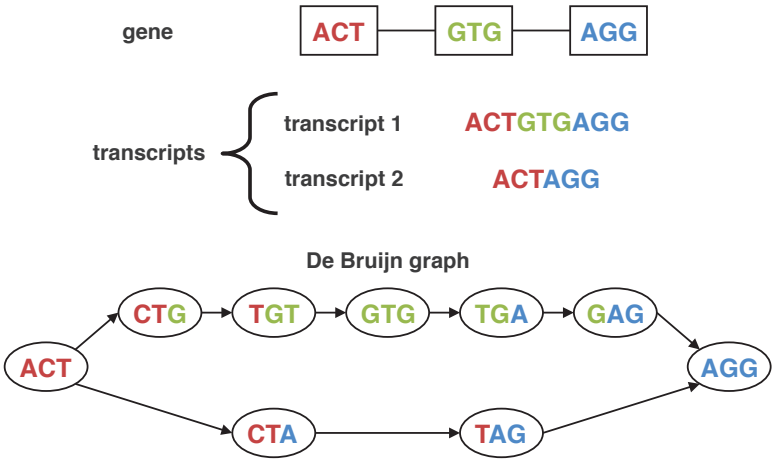

Figure S13. A schematic gene with three exons producing two alternative transcripts. The De Bruijn graph built from the sequences of the transcripts corresponds to a bubble. The upper path spells the skipped exon and its flanking junctions while the lower path spells the junction of the exclusion isoform and has a predictable length.

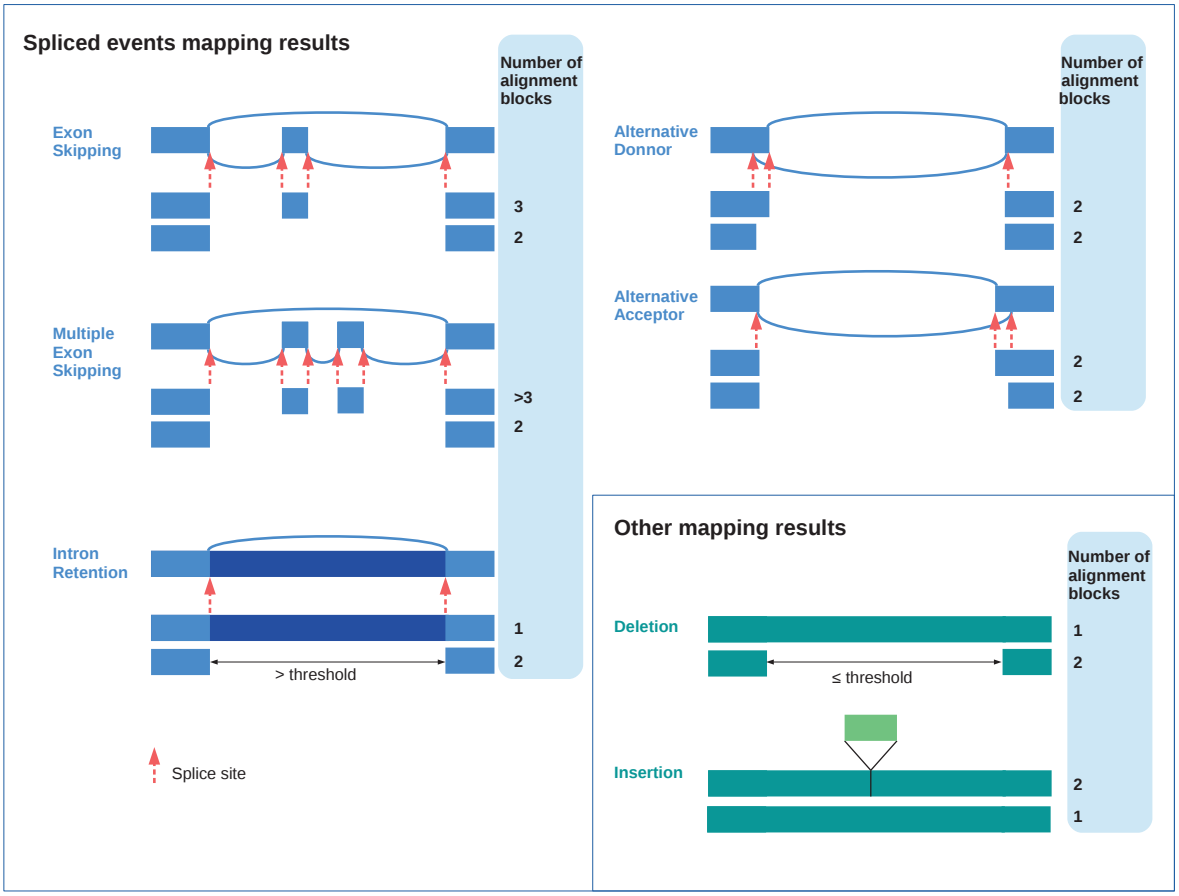

Figure S14. Classification of KISSPLICE events according to the number of blocks in which they map to the reference genome. Paths representing variants of an event are mapped on the reference. Spliced mapping results in blocks, events are then classified by KISSPLICE2REFGENOME according to the block mapping patterns. (Putative) splice sites are noted by SS in red.

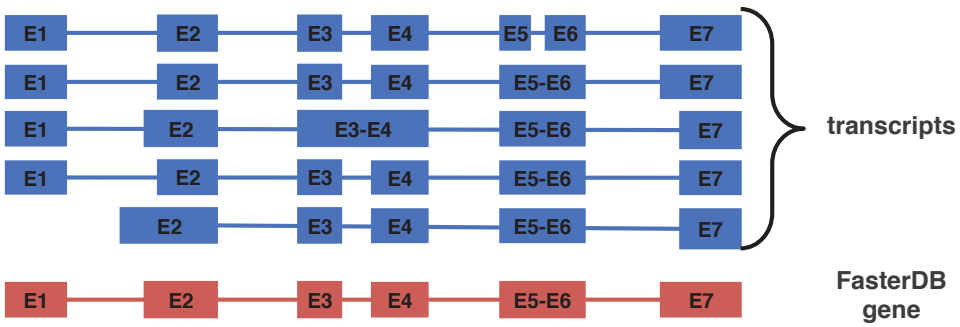

Figure S15. FasterDB exons are defined as the projection of the longer or most frequent exon in the transcripts (except for alternative first or last exons). The whole analysis done with FARLINE is based on these exons.

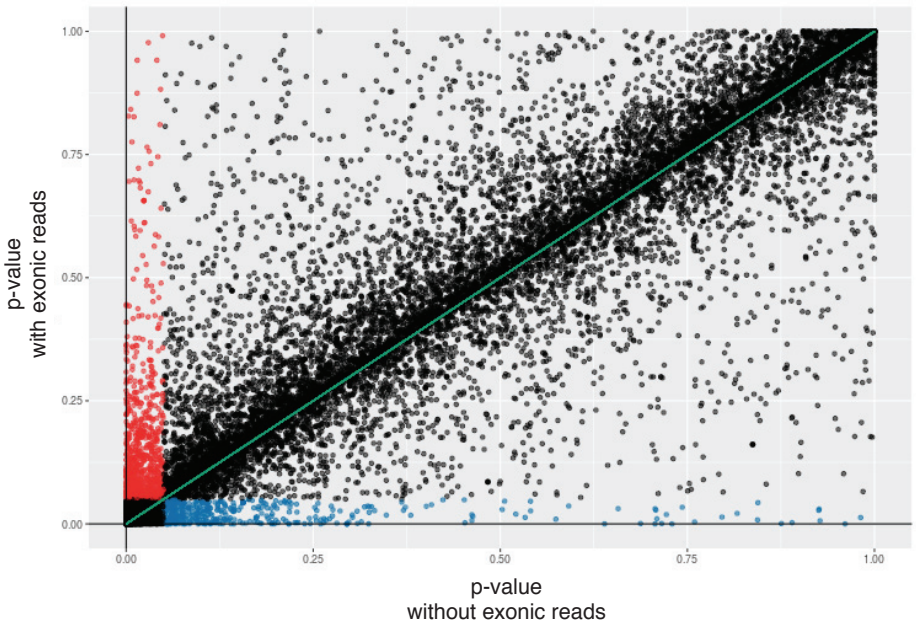

Figure S16. Correlation of the p-values when exonic reads were taken or not into account in the quantification. Red dots and blue dots correspond to ASE predicted to be regulated (p-value$<$0.05) when using junction reads and when using junction and exonic reads respectively.

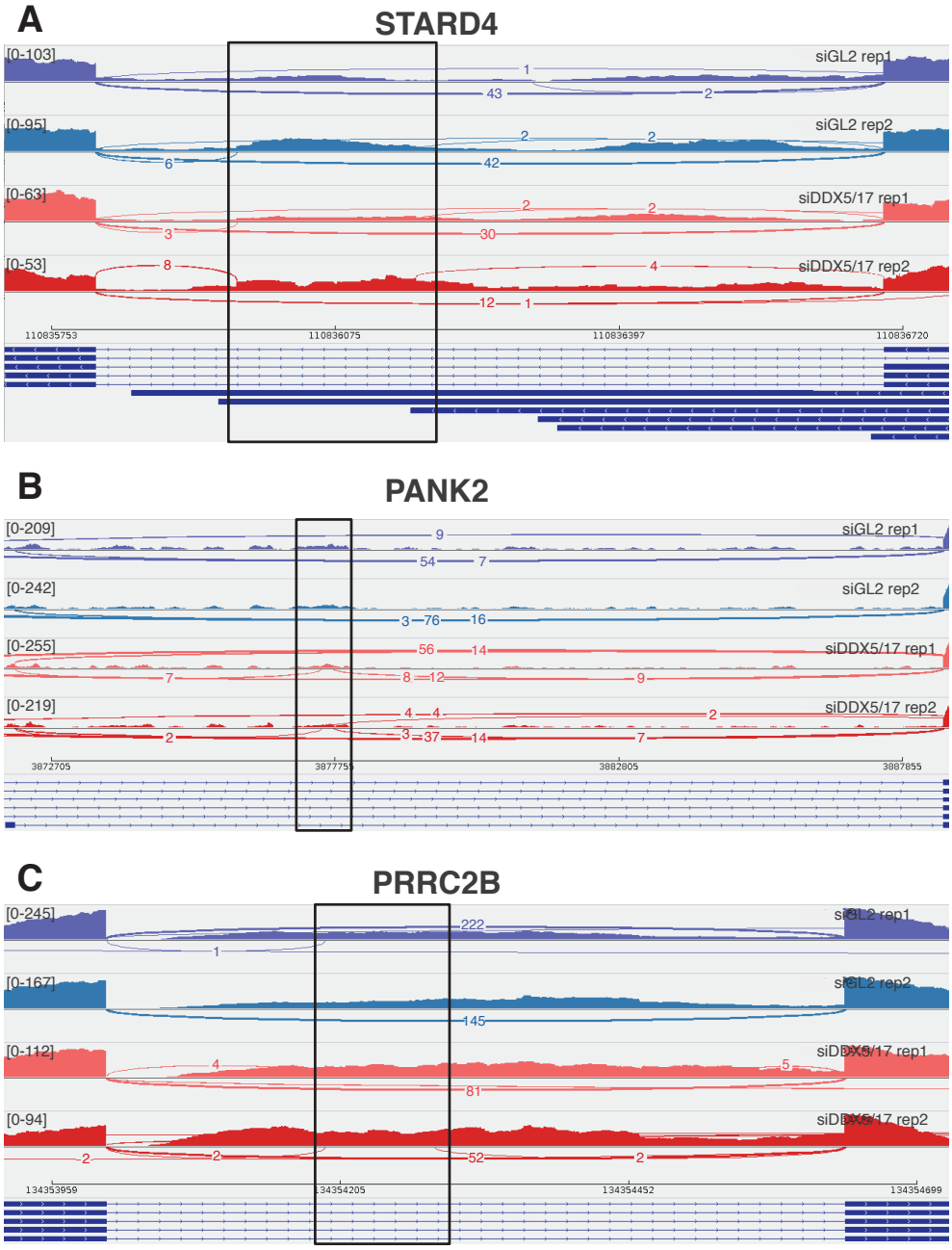

Figure S17. Examples of AS events predicted as differentially spliced between the two conditions in the MCF-7 dataset using junction and exonic reads, but not using only junction reads. A) Exon 6 of *STARD4* is detected as an alternatively skipped exon, but it also overlaps with an alternative last exon. B-C) Exon in intron 3 of *PANK2* gene and exon in intron 18 of *PRRC2B* gene are new exons found by KISSPLICE. These exons are located in poorly spliced introns.

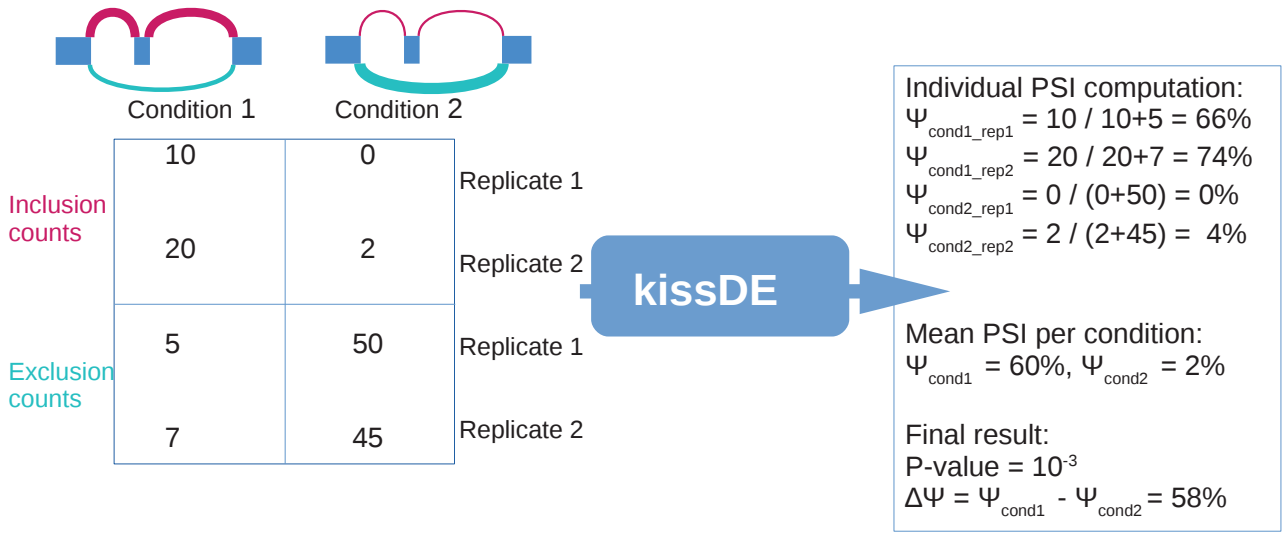

Figure S18. Input and output of the differential analysis. Counts for each replicate of each condition were computed by FARLINE or KISSPLICE. These counts together with the experimental plan are the input of KISSDE. In this example, we show counts for one single event, in practice KISSDE tests all events discovered by one method to spot the differential splicing events. Provided that at least two replicates are available per condition, KISSDE computes p-values and DeltaPSI (ΔΨ) per event, and results are ranked using these two metrics.

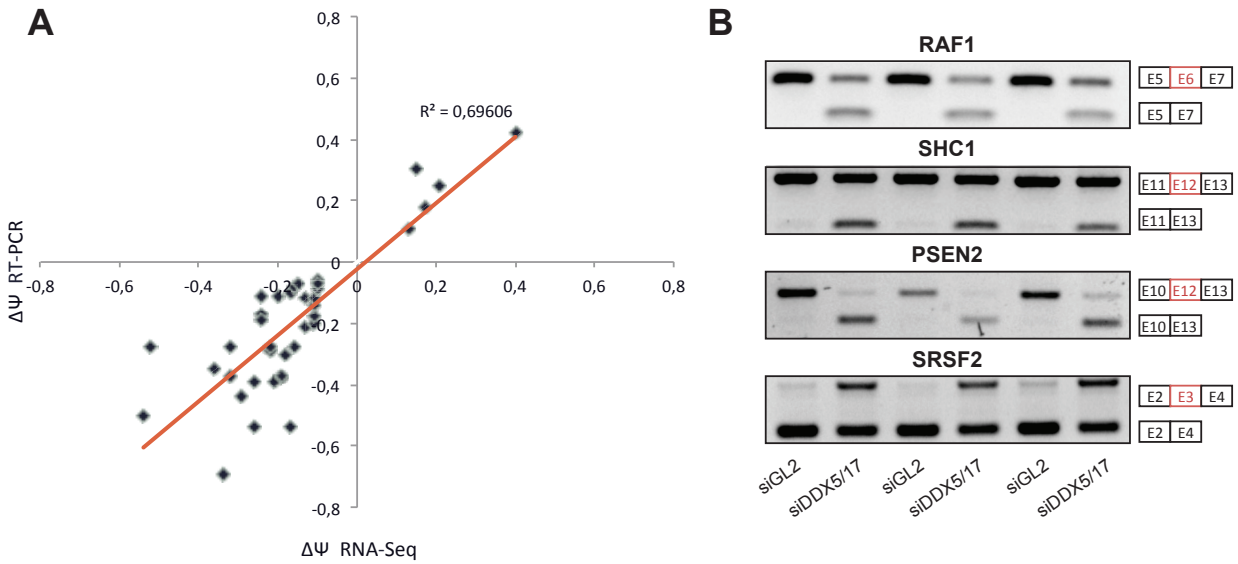

Figure S19. Validations of ASE regulated by the depletion of DDX5 and DDX17 in MCF7 cell line. A) Correlation of the deltaPSI computed from the RNAseq and the deltaPSI computed from the validations by RT-PCR. B) RT-PCR validations of some of the events regulated by the depletion of DDX5 and DDX17 in MCF7 cell line.

B. Discussion

Cette première publication a permis de démontrer l'importance d'utiliser des méthodes mapping- et assembly-first pour être le plus exhaustif possible au niveau de l'identification des événements d'épissage de type saut d'exon (ES). Les forces et faiblesses de ces méthodes sont en effet complémentaires, et il est recommandé d'utiliser l'union de leurs résultats plutôt que leur intersection.

FaRLine (méthode mapping-first, utilisée avec un transcriptome de référence) est ainsi le plus efficace pour détecter des variants rares ou des exons contenant des répétitions. KisSplice (méthode assembly-first), quant à lui, est capable d'identifier des ES dans des gènes récemment dupliqués et des ES d'exons non-annotés, même chez l'Homme où l'annotation est réputée complète. Il semble possible que les avantages de ces méthodes soient conservés dans le contexte d'autres événements d'épissage, comme les donneurs/accepteurs alternatifs ou les rétentions d'intron.

Est-ce que des méthodes mapping-first n'utilisant pas de transcriptome de référence, comme LeafCutter (Li et al., 2016) ou MAJIQ (Vaquero-Garcia et al., 2016), seraient capables de détecter les mêmes jonctions d'épissage non-annotées que KisSplice ? Fondamentalement, la capacité de méthodes mapping-first à détecter ou non de nouvelles jonctions est liée au logiciel d'alignement utilisé en amont : si des jonctions ne sont pas couvertes dans le résultat de cet alignement, elles ne seront détectables par aucun outil d'analyse utilisé par la suite. Puisque certains aligneurs proposent des options pour la détection de telles jonctions (STAR two pass, Hisat2), la détection de formes d'épissage non-annotées devrait être améliorée, bien que cette tâche soit encore très compliquée pour certain reads. Dans ces cas-là, assembler dans un premier temps les reads en contigs plus long, pour ensuite aligner ces contigs, devrait faciliter la tâche des aligneurs et potentiellement mener à la découverte de nouvelles jonctions. Nous avons donc commencé un travail visant à quantifier et caractériser ces instances difficiles, pour identifier clairement les cas où seule une approche assembly-first permet de trouver de nouvelles jonctions d'épissage.

J'ai continué de suivre le développement de K2RG, en ajoutant un script de création d'un fichier de résultat au format GTF pour répondre à des requêtes d'utilisateurs, tout en développant une autre brique logicielle de KisSplice, dédiée à l'analyse différentielle : kissDE.

II. Analyse différentielle des événements d'épissage alternatif

Entre 2012 et 2018, kissDE était un package R (collection de fonctions, données, tests et documentations, écrit dans un format bien défini et partagé avec tous les utilisateurs du logiciel d'analyse statistique R). J'ai contribué au débogage de kissDE, à la création d'un fichier de résultats plus clairs, à la création de nouvelles fonctions du code, à l'amélioration de la méthode de normalisation des comptages, à l'instauration d'une nouvelle méthode d'estimation du paramètre de dispersion nécessaire à la distribution binomiale négative (utilisée pour modéliser les comptages) et au passage de kissDE sous Bioconductor. Pour cette étape, j'ai participé à la réécriture de parties du code pour satisfaire les prérequis du format Bioconductor, ainsi qu'à l'écriture de la vignette et du guide d'utilisation associés à kissDE.

Bioconductor est un projet open source qui vise à redistribuer des packages R dédiés à l'analyse de données de biologie moléculaire. Pour chacun de ces packages, Bioconductor assure une installation facile, une documentation de qualité et une meilleure visibilité auprès de la communauté scientifique concernée.

Le logiciel kissDE est distribué depuis Mai 2019 par Bioconductor. La vignette explicative accompagnant cette distribution est retranscrite dans le paragraphe suivant.

A. Publication (*Bioconductor*)

The 'kissDE' package

Clara Benoit-Pilven [1]***, Camille Marchet*** [2]***, Janice Kielbassa*** [3]***, Audric Cologne*** [1]***, Aurélie Siberchicot*** [1]***, and Vincent Lacroix*** *
[1]

[1]Université de Lyon, Université Lyon 1, CNRS UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France
[2]Univ Rennes, Inria, CNRS, IRISA, France
[3]Synergie Lyon Cancer, Université Lyon 1, Centre Léon Berard, Lyon, France

*vincent.lacroix@univ-lyon1.fr

May 2, 2019

Abstract

kissDE is a package dedicated to the analysis of count data obtained from the quantification of pairs of variants in RNA-Seq data.
It can be used to study splice variants, where the two variants of the pair differ by the inclusion/exclusion of an exonic or intronic region. It can also be used to study genomic variants (whenever they are transcribed), which differ by a single nucleotide variation (SNV) or an indel.
The statistical framework is based on similar hypotheses as *DESeq2* [1] and includes its normalization method using geometric means. Counts are modelled using the negative binomial distribution. We use the framework of the generalised linear model, and we test for association of a variant with a condition using a likelihood ratio test.
This vignette explains how to use this package.
The workflow for SNPs/SNVs is fully described in Lopez-Maestre et al. [2], the workflow for splicing is fully described in Benoit-Pilven et al. [3]

Package

kissDE 1.5.0

Contents

1 Prerequisites

1.1 Use case

kissDE is meant to work on pairs of variants that have been quantified across different conditions. It can deal with single nucleotide variations (SNPs, mutations, RNA editing), indels or alternative splicing.
As *kissDE* was first designed to be a brick of the *KisSplice* [4] pipeline (web page: http://kissplice.prabi.fr/), the `kissplice2counts` function can be directly applied to the output files from *KisSplice* or *KisSplice2refgenome*. Yet, *kissDE* can also run with any other software which produces count data as long as this data is properly formatted.
kissDE was designed to work with at least two replicates for each condition, which means that the minimal input contains the read counts of the variants for 4 different samples, each couple representing a biological condition and its 2 replicates. There can be more replicates and more conditions, but it is not mandatory to have an equal number of replicates in each condition.

1.2 Install and load *kissDE*

In a *R* session, the *BiocManager* package has first to be installed.

```
install.packages("BiocManager")
```

Then, the *kissDE* package can be installed from *Bioconductor* and finally loaded.

```
BiocManager::install("kissDE")
```

```
library(kissDE)
```

1.3 Quick start

Here we present the basic *R* commands for an analysis with *kissDE*. These commands require an external output file of *KisSplice* , for example '`output_kissplice.fa`' (which is not included in this package). To deal with other types of input files, please refer to section 2.1. The funtions used in *kissDE* are `kissplice2counts`, `qualityControl`, `diffExpressedVariants` and `writeOutputKissDE`. For each function, default values of the parameters are used. For more details on functions and their parameters see section 2. Here we assume that there are two conditions ($condition\_1$ and $condition\_2$) with two biological replicates and we also assume that the RNA-Seq libraries are single-end.

```
counts <- kissplice2counts("output_kissplice.fa")
conditions <- c(rep("condition_1", 2), rep("condition_2", 2))
qualityControl(counts, conditions)
results <- diffExpressedVariants(counts, conditions)
writeOutputKissDE(results, output = "kissDE_output.tab")
```

Note that the functions `kissplice2counts` and `diffExpressedVariants` may take some time to run (see section 4.3 for more details on running time).

2 *kissDE*'s workflow

In this section, the successive steps and functions of a differential analysis with *kissDE* are described.

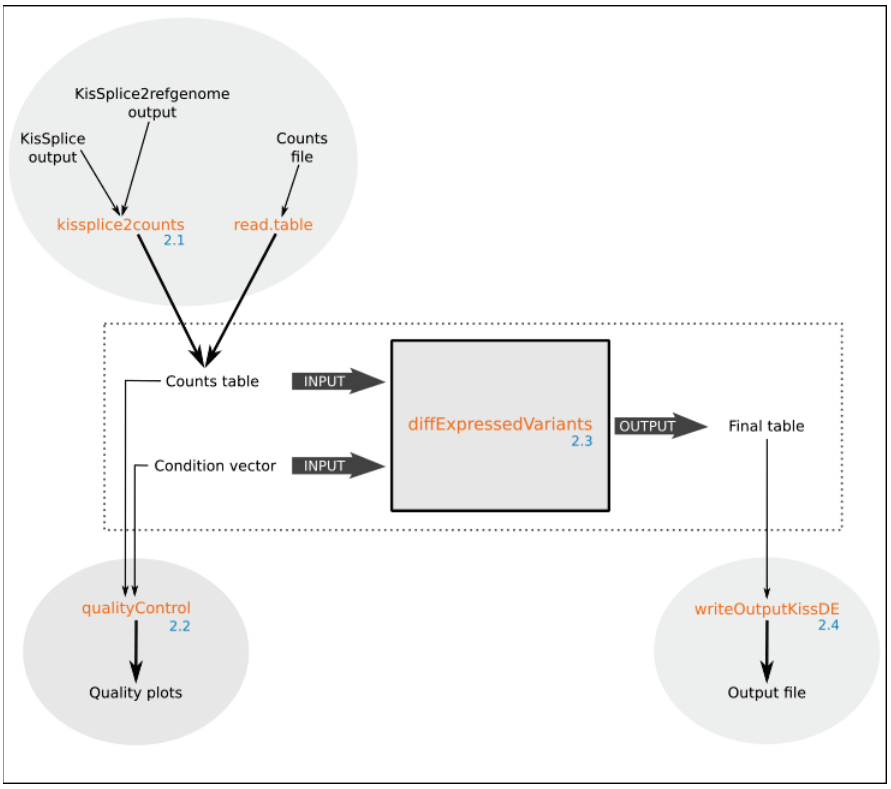

Figure 1: Schema of *kissDE*'s workflow
Numbers in light blue point to the section of this vignette explaining the step.

2.1 Input data

kissDE's input is a table of raw counts and a vector describing the number of conditions and replicates per condition. The table of raw counts can either be directly provided by the user or obtained with *KisSplice* or *KisSplice2refgenome* (http://kissplice.prabi.fr/training/).

2.1.1 Condition vector

The condition vector describes the order of the columns in the count table.
As an example, the counts are ordered as follow: the two first counts represent the two replicates of $condition\_1$ and the two following counts the two replicates of $condition\_2$. In this case, the condition vector for these 2 conditions with 2 replicates per condition, would be:

```
myConditions <- c(rep("condition_1", 2), rep("condition_2", 2))
```

In the case where the input data contains more than 2 conditions, we advise the user to remove samples from the analysis in order to compare 2 conditions only, because *kissDE* was uniquely tested in this context. To remove samples from the analysis the "*" character can be used:

```
myConditions <- c(rep("condition_1", 2), rep("*", 2), rep("condition_3",
    2))
```

Here, there are 3 conditions and 2 replicates per condition, but only $condition\_1$ and $condition\_3$ will be considered in the analysis.
If the count table was loaded from *KisSplice* or *KisSplice2refgenome* output, the condition vector must contain the samples in the same order they were given to *KisSplice* (see sections 2.1.3 and 2.1.4).

Warning: To run kissDE, all conditions must have replicates. So each condition must at least be present twice in the condition vector. If this is not the case, an error message will be printed.

2.1.2 User's own data (without *KisSplice*): table of counts format

Let's assume we work with two conditions ($condition\_1$ and $condition\_2$) and two replicates per condition. An input example table contained in a flat file called '`table_counts_alt_splicing.txt`' is loaded and stored in a `tableCounts` object.
Comment: `fpath1` contains the absolute path of the file on the user's hard disk.

```
fpath1 <- system.file("extdata", "table_counts_alt_splicing.txt",
    package = "kissDE")
tableCounts <- read.table(fpath1, head = TRUE)
```

In *kissDE*, the table of counts must be formatted as follows:

```
head(tableCounts)
```

```
  eventsName eventsLength cond1rep1 cond1rep2 cond2rep1 cond2rep2
1     event1          261       105        41        15        26
2     event1           81         2         5       100       150
3     event2          207        20        17        60        58
4     event2           80        58        33         7         1
5     event3          268        53        26        19        29
6     event3           82         3         1        31        55
```

It must be a data frame with:

- **in rows:**
  - One variation is represented by two lines, one for each variant. For instance, for SNVs, one allele is described in the first line, and the other in the second line. For alternative splicing events, the inclusion isoform and the exclusion isoform have one line each.
  - The header must contain the column names in the flat file.
- **in columns:**
  - The first column (`eventsName`) contains the name of the variation.
  - The second column (`eventsLength`) contains the effective size of the variant in nucleotides (bp). The effective size corresponds to the number of read mapping positions used when estimating the abundance of a variant.
    For the exclusion variant (2nd line), which should correspond to an exon-exon junction, it corresponds to:

$$effectiveLengthExclu = readLength - 2 * overhang + 1 \quad (1)$$

    where $overhang$ corresponds to the minimal number of bases needed to accept that a read is aligned to a junction.
    For the inclusion variant (1st line), it corresponds to:

$$effectiveLengthInclu = effectiveLengthExclu + variablePartLength \quad (2)$$

    where $variablePartLength$ is the length of the region only present in the inclusion variant.
    In the special case where the abundance of the inclusion variant has been estimated using only junction reads, then the effective length of the inclusion variant is:

$$effectiveLengthInclu = 2 * effectiveLengthExclu \quad (3)$$

    This information is used only in the context of alternative splicing. In the context of SNVs, it can be set to 0. It is used to assess which splice variants may induce a frameshift (the difference of length between the inclusion and exclusion variant is not a multiple of 3). It is also used to precisely estimate the PSI (Percent Spliced In).
  - All other columns (`cond1rep1`, `cond1rep2`, `cond2rep1`, `cond2rep2`) contain read counts of a variant in a sample. In the example above, `cond1rep1` is the number of reads supporting this variant in the first replicate of $condition\_1$, `cond1rep2` is the number of reads supporting replicate 2 in $condition\_1$, `cond2rep1` and `cond2rep2` are counts for replicates 1 and 2 of $condition\_2$.

2.1.3 Input table from *KisSplice* output

kissDE was developped to deal with *KisSplice* output, which is in fasta format. Below is the first four lines of an example of *KisSplice* output:

```
headfasta <- system.file("extdata",
    "head_output_kissplice_alt_splicing_fasta.txt", package = "kissDE")
writeLines(readLines(headfasta))
```

```
>bcc_68965|Cycle_4|Type_1|upper_path_length_112|AS1_1|SB1_1|S1_0|ASSB1_0|AS2_0|
SB2_0|S2_0|ASSB2_0|AS3_0|SB3_0|S3_0|ASSB3_0|AS4_1|SB4_0|S4_0|ASSB4_0|AS5_8|SB5_
2|S5_0|ASSB5_1|AS6_13|SB6_4|S6_0|ASSB6_3|AS7_4|SB7_1|S7_0|ASSB7_1|AS8_3|SB8_1|S
8_0|ASSB8_0|rank_0.76503
CACACCAGCCATAAAAAGCGAAAGAATAAAAACCGGCACAGCCCGTCTGGCATGTTTGATTATGACTTTGAGTATGTAT
ATTAGGTTAGGCTGGGAAGTTTTTTTTAAAAAC
>bcc_68965|Cycle_4|Type_1|lower_path_length_82|AB1_21|AB2_12|AB3_12|AB4_2|AB5_5
|AB6_1|AB7_2|AB8_1|rank_0.76503
CACACCAGCCATAAAAAGCGAAAGAATAAAAACCGGCACAGGTATGTATATTAGGTTAGGCTGGGAAGTTTTTTTTAAA
AAC
```

Events are reported in blocks of 4 lines, the first two lines correspond to one variant of the splicing event (or one allele of the SNV), the following two lines correspond to the other variant (or the other allele). As for all fasta file, there is a header line beginning with the `>` symbol and a line with the sequence. Each variant correspond to one entry in the fasta file.

Headers contain information used in *kissDE*. In the example, there are:

- elements shared by the headers of the two variants:
  - `bcc_68965|Cycle_4` is the event's ID.
  - `Type_1` means that the sequences correspond to a splicing event. `Type_0` corresponds to SNVs.
- elements that are specific to a variant:
  - `upper_path_length_112` and `lower_path_length_82` gives the length of the nucleotide sequences. Upper path and lower path are a denomination for the representation of each variant in *KisSplice*'s graph. For alternative splicing events, the upper path represents the inclusion isoform and the lower path the exclusion isoform.
  - `AS1_1|SB1_1|S1_0|ASSB1_0|AS2_0|SB2_0|S2_0|ASSB2_0|AS3_0| SB3_0|...` and `AB1_21|AB2_12|AB3_12|AB4_2|AB5_5|...` summarizes the counts found by *KisSplice* quantification step. Here *KisSplice* was run with the option `counts` set to 2. For the upper path, we have 4 counts for each sample: AS, SB, S and ASSB. For the lower path, we have 1 count per sample: AB. The different reads categories are shown on Figure 2. There are 8 sets of counts because we gave 8 files in input to *KisSplice* (denotated by the number before the "_" character). Each count (denotated by the number after the "_" character) corresponds to the reads coming from each file that could be mapped on the variant, in the order they have been passed to *KisSplice*.
- a rank information which is a deprecated measure.

kissDE can be used on any type of events output by *KisSplice* (0: SNV, 1: alternative splicing events, 3: indels,...). The user should refer to *KisSplice* manual (http://kissplice.prabi.fr/documentation/) for further questions about the *KisSplice* format and its output.

To be used in *kissDE*, *KisSplice* output must be converted into a table of counts. This can be done with the `kissplice2counts` function. In the example below, the *KisSplice* output file called '`output_kissplice_alt_splicing.fa`', included in the *kissDE* package, is loaded. The table of counts yielded by the `kissplice2counts` function is stored in `myCounts`.
Comment: `fpath2` contains the absolute path of the file on the user's hard disk.

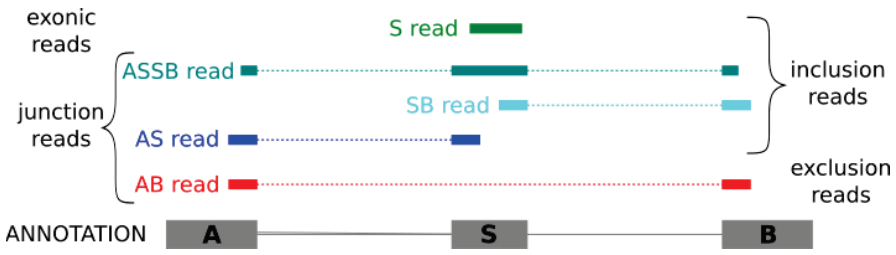

Figure 2: Different categories of reads
In this figure, we show an example of an alternative skipped exon. AS reads correspond to reads spanning the junction between the excluded sequence and its left flanking exon, SB to reads spanning the junction between the excluded sequence and its right flanking exon, ASSB to reads spanning the two inclusion junctions, S to reads entirely included in the alternative sequence and AB to reads spanning the junction between the two flanking exons. S reads correspond to exonic reads and all other categories of reads represented here correspond to junction reads.

```
fpath2 <- system.file("extdata", "output_kissplice_alt_splicing.fa",
    package = "kissDE")
myCounts <- kissplice2counts(fpath2, counts = 2, pairedEnd = TRUE)
```

The counts returned by `kissplice2counts` are extracted from the *KisSplice* header. By default, `kissplice2counts` expects single-end reads and one count for each variant.
The `counts` parameter of `kissplice2counts` must be the same as the `counts` parameter used to obtain data with *KisSplice*. The possible values are 0, 1 or 2. 0 is the default value for both `kissplice2counts` and *KisSplice*.
The user can also specify the `pairedEnd` parameter in `kissplice2counts`. If RNA-Seq libraries are paired-end, `pairedEnd` should be set to `TRUE`. In this case, the `kissplice2counts` function expects the counts of the paired-end reads to be next to each other. If it is not the case, an additional `order` parameter should be used to indicate the actual order of the counts. For instance, if the experimental design is composed of two conditions with two paired-end replicates and if the input in *KisSplice* followed this order:
cond1_sample1_readpair1, cond1_sample2_readpair1, cond2_sample1_readpair1, cond2_sample2_readpair1, cond1_sample1_readpair2, cond1_sample2_readpair2, cond2_sample1_readpair2 and cond2_sample2_readpair2.
The order vector should be equal to `c(1,2,3,4,1,2,3,4)`.
An example of a paired-end dataset run with `counts` equal to 0 is shown in section 4.2.

`kissplice2counts` returns a list of four elements, including `countsEvents` which contains the table of counts required in *kissDE*.

```
names(myCounts)
```

```
[1] "countsEvents"    "psiInfo"         "exonicReadsInfo" "k2rgFile"
```

```
head(myCounts$countsEvents)
```

```
       events.names events.length counts1 counts2 counts3 counts4
1  bcc_68965|Cycle_4           112       2       1      23       8
2  bcc_68965|Cycle_4            82      33      14       6       3
3  bcc_83285|Cycle_2           180     108      47      33      36
4  bcc_83285|Cycle_2            81       2       5     100     150
5 bcc_161433|Cycle_2           127      20      17      60      58
6 bcc_161433|Cycle_2            80      58      33       7       1
```

`myCounts$countsEvents` has the same structure as the `tableCounts` object in the section 2.1.2. It is a data frame with:

- **in rows:** One variation is represented by two lines, one for each variant. For instance for SNVs, one allele is described in the first line and the other in the second line. For alternative splicing events (as in this example), the inclusion and the exclusion isoform have one line each.
- **in columns:**
  - The first column (`events.names`) contains the name of the variation, using *KisSplice* notation.
  - The second column (`events.length`) contains the size of the variant in bp, extracted from the *KisSplice* header.
  - All others columns (`counts1, counts2, counts3, counts4`) contain counts for each replicate in each condition for the variant.

2.1.4 Input table from *KisSplice2refgenome* output

The `kissplice2counts` function can also deal with *KisSplice2refgenome* output data, in this case the `k2rg` parameter has to be set to `TRUE`. *KisSplice2refgenome* allows the annotation of the alternative splicing events. It assigns each event a gene and a type of alternative splicing event, among which: Exon Skipping (ES), Intron Retention (IR), Alternative Donor (AltD), Alternative Acceptor (AltA). Interested users should refer to *KisSplice2refgenome* manual for further questions about *KisSplice2refgenome* format and output (http://kissplice.prabi.fr/tools/kiss2refgenome/).

In the example below, '`output_k2rg_alt_splicing.txt`', a *KisSplice2refgenome*'s output included in the *kissDE* package, is loaded. The `kissplice2counts` function uses the same `counts` and `pairedEnd` parameters as explained in the section 2.1.3. The table of counts yielded by the `kissplice2counts` function is stored in `myCounts_k2rg`. It has exactly the same structure as detailed in section 2.1.3.
Comment: `fpath3` contains the absolute path of the file on the user's hard disk.

```
fpath3 <- system.file("extdata", "output_k2rg_alt_splicing.txt",
    package = "kissDE")
myCounts_k2rg <- kissplice2counts(fpath3, counts = 2, pairedEnd = TRUE,
    k2rg = TRUE)
names(myCounts_k2rg)
```

```
[1] "countsEvents"    "psiInfo"         "exonicReadsInfo" "k2rgFile"
```

```
head(myCounts_k2rg$countsEvents)
```

```
       events.names events.length counts1 counts2 counts3 counts4
1  bcc_68965|Cycle_4           112       2       1      23       8
2  bcc_68965|Cycle_4            82      33      14       6       3
3  bcc_83285|Cycle_2           180     108      47      33      36
4  bcc_83285|Cycle_2            81       2       5     100     150
5 bcc_161433|Cycle_2           127      20      17      60      58
6 bcc_161433|Cycle_2            80      58      33       7       1
```

The *KisSplice2refgenome* output contains information about the type of splicing events. By default, all of the splicing events are analysed in *kissDE*, but it is also possible to focus on subtypes of events. This events selection will speed up *kissDE*'s running time and improve statistical power for choosen events. To do this, the `kissplice2counts` function contains two parameters: `keep` and `remove`. Both take a character vector indicating the types of events to keep or remove. The event names must be part of this list: `deletion`, `insertion`, `IR`, `ES`, `altA`, `altD`, `altAD`, `alt`, `unclassified`.
Thus, if the user is only interested in intron retention events, the `keep` option should be set to `c("IR")`. If the user isn't interessed in deletions and insertions, the `remove` option should be equal to `c("insertion", "deletion")`.
The `keep` and `remove` parameters can be used at the same time only if `ES` is part of the `keep` vector. The `remove` vector will then act on the different types of exon skipping: multi-exon skipping (`MULTI`) or exon skipping associated with an alternative acceptor site (`altA`), an alternative donor site (`altD`), both alternative acceptor and donor site (`altAD`) or an undetermined alternative splice site (`alt`). Thus, in this specific case, the `remove` vector should contain names from this list: `MULTI`, `altA`, `altD`, `altAD`, `alt`.

If the user wants to analyse only cassette exon events (i.e., a single exon is skipped or included), the following command should be used:

```
myCounts_k2rg_ES <- kissplice2counts(fpath3, counts = 2, pairedEnd = TRUE,
    k2rg = TRUE, keep = c("ES"), remove = c("MULTI", "altA",
        "altD", "altAD", "alt"))
```

2.2 Quality Control

kissDE contains a function that allows the user to control the quality of the data and to check if no error occured at the data loading step. This data quality assessment is essential and should be done before the differential analysis.

The `qualityControl` function takes as input a count table (see sections 2.1.2, 2.1.3 and 2.1.4) and a condition vector (see section 2.1.1):

```
qualityControl(myCounts, myConditions)
```

It produces 2 graphs:

- a heatmap of the sample-to-sample distances using the 500 most variant events (see left panel of Figure 3)
- the factor map formed by the first two axes of a principal component analysis (PCA) using the 500 most variant events (see right panel of Figure 3)

These two graphs show the similarities and the differences between the analyzed samples. Replicates of the same condition are expected to cluster together. If this is not the case, the user should check if the order of the samples in the count table and in the condition vector is the same. If it is, this could mean that a sample is contaminated or has an abnormality that will influence the differential analysis. The user can then go back to the quality control of the raw data to solve the problem or decide to remove the sample from the analysis.

In the heatmap plot, the samples that cluster together are from the same condition. In the PCA plot, the first principal component (PC1) summarize 90.2% of the total variance of the dataset. This first axis clearly separates the 2 conditions.

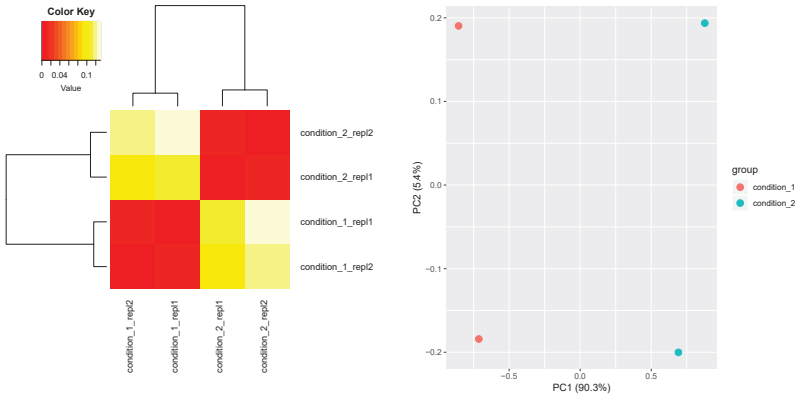

Figure 3: Quality control plots
Left: Heatmap of the sample-to-sample distances. Right: Principal Component Analysis.

The created graphs can be saved by setting the `storeFigs` parameter of the `qualityControl` function to `TRUE` (then graphs are stored in a '`kissDEFigures`' folder, created in a temporary directory, which is removed at the end of the user *R* session) or to the path where the user wants to store his/her graphs. We recommend to use this parameter when the `qualityControl` function is used in an automatized workflow.

To customize the PCA plot, the data frame used for this plot can be extracted by setting the option `returnPCAdata` to `TRUE` as follows:

```
PCAdata <- qualityControl(myCounts, myConditions, returnPCAdata = TRUE)
```

2.3 Differential analysis

When data are loaded, the differential analysis can be run using the `diffExpressedVariants` function. This function has two mandatory parameters: a count table (`countsData` parameter, see sections 2.1.2, 2.1.3 and 2.1.4) and a condition vector (`conditions` parameter, see section 2.1.1).

In the example below, the differential analysis results are stored in the `myResults` object:

```
myResults <- diffExpressedVariants(countsData = myCounts,
    conditions = myConditions)
```

The `diffExpressedVariants` function has three parameters to change the filters or the flags applied on the data, one parameter to indicate if the replicates are technical or biological, and one parameter to indicate how many cores should be used :

- `pvalue`: By default, the p-value threshold to output the significant events is set to 1. So all variants are output in the final table. This parameter must be a numeric value between 0 and 1. Be aware that by setting `pvalue` to 0.05, only events that have been identified as significant between the conditions with a false discovery rate (FDR) $\leqslant$ 5% will be present in the final table. A posteriori changing this threshold will require to re-run the differential analysis.
- `filterLowCountsVariants`: This parameter allows to change the threshold to filter low expressed events before testing (as explained in section 3.3). By default, it is set to 10.

- `flagLowCountsConditions`: This parameter allows to change the threshold to flag low expressed events (as explained in section 3.6). By default, it is set to 10.
- `technicalReplicates`: Boolean value indicating if the user is working with technical replicates only (we do not advise users to mix biological and technical replicates in their analyses). If this parameter is set to `TRUE`, the counts will be modeled with a Poisson distribution. If it is equal to `FALSE`, the counts will be modeled with a Negative Binomial distribution. For more information, see section 3.2. By default, this option is set to `FALSE`.
- `nbCore`: An integer value indicating how many cores should be used for the computation. This parameter should be strictly lower than the number of core of the computer (`nbCore` $<$ nbr computer cores -1). By default, this parameter is set to 1, meaning that the computation are not parallelized.

The `diffExpressedVariants` function returns a list of 6 objects:

```
 names(myResults)

[1] "finalTable"          "correctedPVal"       "uncorrectedPVal"
[4] "resultFitNBglmModel" "f/psiTable"          "k2rgFile"
```

The `uncorrectedPVal` and `correctedPval` outputs are numeric vectors with p-values before and after correction for multiple testing. `resultFitNBglmModel` is a data frame containing the results of the fitting of the model to the data. `k2rgFile` is a string containing either the *KisSplice2refgenome* file path and name or NULL if no *KisSplice2refgenome* file was used as input. For explanations about the `finalTable` and `f/psiTable` outputs, see section 2.4.1 and section 2.4.2, respectively.

To visualize the distribution of the p-values before the application of the Benjamini-Hochberg [5] multiple testing correction procedure, the histogram of the p-values before correction can be plotted by using the following command:

```
 hist(myResults$uncorrectedPVal, main = "Histogram of p-values",
     xlab = "p-values", breaks = 50)
```

Because the dataset used here is small ($\sim$ 100 lines), the histograms of the two complete datasets presented in the case studies (section 4) are represented. As expected, the histograms show a uniform distribution with a peak near 0 (Figure 4).

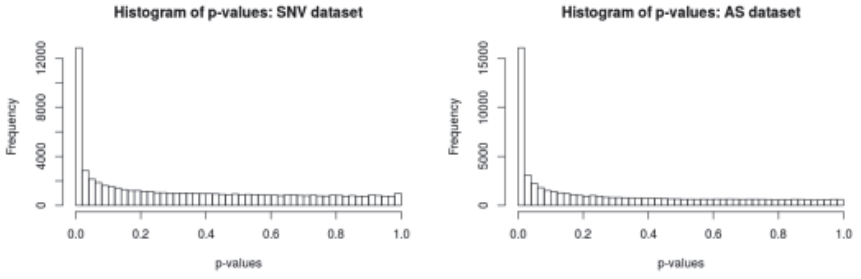

Figure 4: Distribution of p-values before correction for multiple testing
Left: for the complete dataset presented in section 4.2. Right: for the complete dataset presented in section 4.1.

2.4 Output results

2.4.1 Final table

The `finalTable` object is the main output of the `diffExpressedVariants` function. The first 3 rows of the `myResults$finalTable` output are as follows:

```
print(head(myResults$finalTable, n = 3), row.names = FALSE)
```

```
                    ID Length_diff Variant1_condition_1_repl1_Norm
      bcc_83285|Cycle_2          99                              86
      bcc_68965|Cycle_4          30                               2
bcc_135201|Cycle_433392         104                              80
Variant1_condition_1_repl2_Norm Variant1_condition_2_repl1_Norm
                             50                              36
                              1                              25
                             56                              31
Variant1_condition_2_repl2_Norm Variant2_condition_1_repl1_Norm
                             38                               2
                              8                              26
                             35                               2
Variant2_condition_1_repl2_Norm Variant2_condition_2_repl1_Norm
                              5                             108
                             15                               6
                              1                              31
Variant2_condition_2_repl2_Norm Adjusted_pvalue Deltaf/DeltaPSI lowcounts
                            158        0.00e+00          -0.770     FALSE
                              3        7.93e-10           0.703     FALSE
                             58        0.00e+00          -0.696     FALSE
```

The columns of this table contain the following information:

- `ID` is the event identifier. Each event is represented by one row in the table.
- `Length_diff` contains the variable part length in a splicing event. It is the length difference between the upper and lower path. This column is not relevant for SNVs.
- `Variant1_condition_1_repl1_Norm` and following columns contain the counts for each replicate of each variant after normalization (raw counts are normalized as in the *DESeq2 Bioconductor R* package, see details in section 3.1). The first half of these columns concerns the first variant of each event, the second half the second variant.
- `Adjusted_pvalue` contains p-values adjusted by a Benjamini-Hochberg procedure.
- `Deltaf/DeltaPSI` summarizes the magnitude of the effect (see details in section 3.7).
- `lowcounts` contains booleans which flag low counts events as described in section 3.6. A `TRUE` value means that the event has low counts (counts below the chosen threshold).

In the `finalTable` output, events are sorted by p-values and then by magnitude of effect (based on their absolute values), so that the top candidates for further investigation/validation appear at the beginning of the output.

Warning: When the p-value computed by kissDE is lower than the smallest number greater than zero that can be stored (i.e., 2.2e-16), this p-value is set to 0.

To save results, a tab-delimited file can be written with `writeOutputKissDE` function where an `output` parameter (containing the name of the saved file) is required. Here, the `myResults` output is saved in a file called 'results_table.tab':

```
writeOutputKissDE(myResults, output = "kissDE_results_table.tab")
```

Users can choose to export only events passing some thresholds on adjusted p-value and/or Deltaf/DeltaPSI using the options `adjPvalMax` and `dPSImin` of the `writeOutputKissDE` function. For example, if we want to save in a file called 'results_table_filtered.tab' only events with the adjusted p-value $\leqslant 0.05$ and the Deltaf/DeltaPSI absolute value $\geqslant 0.10$, the following command can be used:

```
writeOutputKissDE(myResults, output = "kissDE_results_table_filtered.tab",
    adjPvalMax = 0.05, dPSImin = 0.1)
```

If the counts table was built from a *KisSplice2refgenome* output with the `kissplice2counts` function, running the `writeOutputKissDE` will write a file merging results of differential analysis with *KisSplice2refgenome* data. As previously explained (section 2.4.1), users can choose to save only events passing thresholds:

```
writeOutputKissDE(myResults_K2RG, output = "kissDE_K2RG_results_table.tab",
    adjPvalMax = 0.05, dPSImin = 0.1)
```

2.4.2 f/PSI table

The `f/psiTable` output of the `diffExpressedVariants` function contains the f values for SNV analysis or PSI values for alternative splicing analysis (see details and computation in section 3.7) for each event in each sample. The first three rows of the `f/psiTable` output of the `myResults` object (created in the section 2.3) look like this:

```
                  ID condition_1_repl1 condition_1_repl2 condition_2_repl1
1 bcc_100903|Cycle_0           0.00984            0.0195           0.00607
2 bcc_108176|Cycle_0           0.03805            0.0614           0.03844
3 bcc_120508|Cycle_0           0.94526            0.9477           0.96531
  condition_2_repl2
1            0.0119
2            0.0296
3            0.9414
```

This output can be useful to carry out downstream analysis or to produce specific plots (like heatmap on f/PSI events). To use this information with external tools, this table can be saved in a tab-delimited file (here called 'result_PSI.tab'), setting the `writePSI` parameter to `TRUE` in the `writeOutputKissDE` function:

```
writeOutputKissDE(myResults, output = "result_PSI.tab", writePSI = TRUE)
```

3 *kissDE*'s theory

In this section, the different steps of the *kissDE* main function, `diffExpressedVariants`, are detailed. They are summarized in the Figure 5.

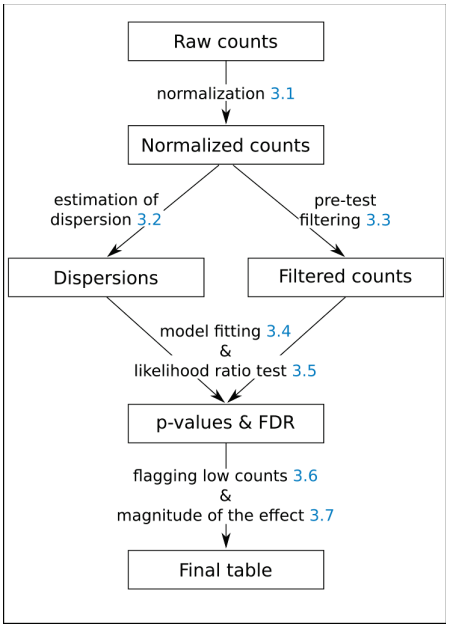

Figure 5: The different steps of the `diffExpressedVariants` function
Numbers in light blue point to the section of this vignette explaining the step.

3.1 Normalization

In a first step, counts are normalized with the default normalization methods provided by the *DESeq2* [1] package. The size factors are estimated using the sum of counts of both variants for each event, which is a proxy of the gene expression. By using this normalization, we correct for library size, because the sequencing depth can vary between samples.

3.2 Estimation of dispersion

A model to describe the counts distribution is first chosen. When working with technical replicates (`technicalReplicates = TRUE` in `diffExpressedVariants`), the Poisson model (model $\mathcal{M}(\phi = 0)$) is chosen in *kissDE*.
When working with biological replicates (`technicalReplicates = FALSE` in `diffExpressedVariants`), the Poisson distribution's variance parameter is in general not flexible enough to describe the data, because replicates add several sources of variance.
This overdispersion is often modeled using a Negative Binomial distribution. In *kissDE*, the overdispersion parameter, ϕ, is estimated using the *DSS* *R* package [6, 7, 8, 9] (model $\mathcal{M}(\phi = \phi^i_{\text{DSS}})$.
The *DSS* package (and, to our knowledge, every other package estimating the overdispersion of the Negative Binomial model) is suited for differential expression analysis (one count per sample). In differential splicing and SNV analysis, two counts (one for each splice variant or allele) are associated with each sample. In order to mimic gene expression, the overdispersion parameter ϕ is estimated on the sum of the splice variant or allele counts of each sample.

3.3 Pre-test filtering

If global counts for both variants are too low (option `filterLowCountsVariants`), the event is not tested. The rationale behind this filter is to speed up the analysis and gain statistical power.
Here we present an example to explain how `filterLowCountsVariants` option works. Let's assume that there are two conditions and two replicates per condition. `filterLowCountsVariants` keeps its default value, 10.

	Condition 1		Condition 2		Sum by variant
	replicate 1	replicate 2	replicate 1	replicate 2	
Variant 1	2	1	3	2	$2+1+3+2=8 < 10$
Variant 2	8	0	1	0	$8+0+1+0=9 < 10$

Table 1: Example of an event filtered out before the differential analysis, because less than 10 reads support each variant

In this example (Table 1), the two variants have global counts less than 10, this event will be used to compute the overdispersion, but will not be used to compute the models. It will neither appear in the result table.

3.4 Model fitting

Then we design two models to take into account interactions with variants (SNVs or alternative isoforms) and experimental conditions as main effects. We use the generalised linear model framework. The expected intensity λ_{ijk} can be written as follows:

$$\mathcal{M}_I : \log \lambda_{ijk} = \mu + \alpha_i + \beta_j \quad (4)$$

$$\mathcal{M}_\infty : \log \lambda_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad (5)$$

where μ is the local mean expression of the transcript that contains the variant, α_i the effect of variant i on the expression, β_j the contribution of condition j to the total expression, and $(\alpha\beta)_{ij}$ the interaction term.

To avoid singular hessian matrices while fitting models, pseudo-counts (*i.e.*, systematic random allocation of ones) were considered for variants showing many zero counts.

3.5 Likelihood ratio test

To select between $\mathcal{M}_I$ and $\mathcal{M}_\infty$, we perform a Likelihood Ratio Test (LRT) with one degree of freedom. In the null hypothesis $H_0 : \{(\alpha\beta)_{ij} = 0\}$, there is no interaction between variant and condition. For events where H_0 is rejected, the interaction term is significant to explain the count's distribution, which leads to conclude to a differential usage of a variant across conditions. p-values are then adjusted with a 5% false discovery rate (FDR) following a Benjamini-Hochberg procedure [5] to account for multiple testing.

3.6 Flagging low counts

If in at least $n-1$ conditions (be n the number of conditions ≥ 2) an event has low counts (option `flagLowCountsConditions`), it is flagged (`TRUE` in the last column of the `finalTable` output).
In the example Table 2, we can see that the counts are quite contrasted, variant 1 seemed more expressed in condition 2 and variant 2 in condition 1. Moreover, this event has enough counts for each variant not to be filtered out when the `filterLowCountsVariants` parameter is set to 10:

	Condition 1		Condition 2		Sum by variant
	replicate 1	replicate 2	replicate 1	replicate 2	
Variant 1	1	0	60	70	1+0+60+70=131 $>$ 10
Variant 2	5	3	10	20	5+3+10+20=38 $>$ 10
Sum by condition	**9** $<$ 10		160 $>$ 10		

Table 2: Example of an event flagged as having low counts, because less than 10 reads support this event in the first condition

However, in $n-1$ (here 1) condition, the global count for one condition is less than 10 (9 for condition 1), so `flagLowCountsConditions` option will flag this event as `'Low_Counts'`. This event may be interesting because it has the potential to be found as differential. However, it will be hard to validate it experimentally, because the gene is poorly expressed in condition 1.

3.7 Magnitude of the effect

When a gene is found to be differentially spliced between two conditions, or an allele is found to be differentially present in two populations/conditions, one concern which remains is to quantify the magnitude of this effect. Indeed, especially in RNA-Seq, where some genes are very highly expressed (and hence have very high read counts), it is often the case that we detect significant (p-value $\leqslant 0.05$) but weak effects.

When dealing with genomic variants, we quantify the magnitude of the effect using the difference of allele frequencies (f) between the two conditions. When dealing with splicing variants, we quantify the magnitude of the effect using the difference of Percent Spliced In (PSI) between the two conditions. These two measures turn out to be equivalent and can be summarized using the following formula:

$$PSI = f = \frac{\#counts * \_variant_1}{\#counts * \_variant_1 + \#counts\_variant_2} \quad (6)$$

$$\Delta PSI = PSI_{cond1} - PSI_{cond2} \quad (7)$$

$$\Delta f = f_{cond1} - f_{cond2} \quad (8)$$

In this formula, $\#counts * \_variant_1$ correspond to the normalized number of reads of the $variant_1$, itself normalized for the variant length. Indeed, by construction, $variant_1$ always have a length greater than or equal to the $variant_2$. That's why we divide the normalized number of reads of the $variant_1$ by the ratio of the length of the $variant_1$ and the $variant_2$.

The ΔPSI/Δf is computed as follows:

- First, individual (per replicate) PSI/f are calculated. If counts for both upper and lower paths are too low (< 10) after normalization, the individual PSI/f are not computed.
- Then mean PSI/f are computed for each condition. If more than half of the individual PSI/f were not calculated at the previous step, the mean PSI/f is not computed either.
- Finally, we output ΔPSI/Δf. Unless one of the mean PSI/f of a condition could not be computed, ΔPSI/Δf is calculated subtracting one condition PSI/f from another. ΔPSI/Δf absolute value vary between 0 and 1, with values close to 0 indicating low effects and values close to 1 strong effects. Note that the conditions are ordered alphabetically, and that *kissDE* substract the condition coming first in the alphabet to the other.

4 Case studies

To detect SNVs (SNPs, mutations, RNA editing) or alternative splicing (AS) in the expressed regions of the genome, *KisSplice* can be run on RNA-seq data. Counts can then be analysed using *kissDE*. We present two distinct case study with *kissDE*: analysis of AS events and analysis of SNVs.

4.1 Application of *kissDE* to alternative splicing

This first example corresponds to the case of differential analysis of alternative splicing (AS) events. The sample data presented here is a subset of the case study used in [3] (http://kissplice.prabi.fr/pipeline_ks_farline/).

4.1.1 Dataset

The data used in this example comes from the ENCODE project [10]. The samples are from a neuroblastoma cell line, SK-N-SH, with or without a retinoic acid treatment. Each condition is composed of two biological replicates. The data are paired-end.

In a preliminary step, *KisSplice* has been run to analyse these two conditions. Results from *KisSplice* (type 1 events) were then mapped to the reference genome with *STAR* [11] and analyzed with *KisSplice2refgenome*. *KisSplice2refgenome* enables to annotate the AS events discovered by *KisSplice*. It assigns to each event a gene and a type of alternative splicing (Exon Skipping (ES), Intron Retention (IR), Alternative Donor (AltD), Alternative Acceptor (AltA), ...).
For further information on these tools (*KisSplice* and *KisSplice2refgenome*), please refer to the manual that can be found on this web page: http://kissplice.prabi.fr/.

The output file of *KisSplice2refgenome* is a tab-delimited file that stores the annotated alternative splicing events found in the dataset. Below is an extract of this file (the first 3 rows and first 10 columns), where each row is one alternative splicing event of our data:

```
        Gene_Id Gene_name Chromosome_and_genomic_position Strand Event_type
ENSG00000163875     MEAF6          chr1:37962165-37967445      -         ES
ENSG00000117620   SLC35A3        chr1:100435679-100459133      +         ES
ENSG00000125814      NAPB         chr20:23375598-23383670      -         ES
Variable_part_length Frameshift_? CDS_?   Gene_biotype
                  30           No   Yes protein_coding
                  99           No   Yes protein_coding
                  47          Yes   Yes protein_coding
number_of_known_splice_sites/number_of_SNPs
              all_splice_sites_known_(4_ss)
              all_splice_sites_known_(4_ss)
              all_splice_sites_known_(4_ss)
```

4.1.2 Load data

The `kissplice2counts` function allows to load directly the *KisSplice2refgenome* output file (here called 'output_k2rg_alt_splicing.txt') into a format compatible with *kissDE*'s main functions.
Comment: `fileInAS` contains the absolute path of the file on the user's hard disk.
The `k2rg` parameter is set to `TRUE` to indicate that the file comes from *KisSplice2refgenome* and not directly from *KisSplice*. As these samples are paired-end, the `pairedEnd` parameter is set to `TRUE`. The `counts` parameter must be set to the same value (i.e., 2) used in *KisSplice* and *KisSplice2refgenome* to indicate which type of counts are given in the input. Here the exonic reads are not taken into account (`exonicReads = FALSE`). Only junction reads will be used (see Figure 2).

The table of counts is stored in a `myCounts_AS` object (for a detailed description of its structure, see section 2.1.4):

```
fileInAS <- system.file("extdata", "output_k2rg_alt_splicing.txt",
    package = "kissDE")
myCounts_AS <- kissplice2counts(fileInAS, pairedEnd = TRUE, k2rg = TRUE,
    counts = 2, exonicReads = FALSE)
head(myCounts_AS$countsEvents)
```

```
       events.names events.length counts1 counts2 counts3 counts4
1  bcc_68965|Cycle_4           112       2       1      23       8
2  bcc_68965|Cycle_4            82      33      14       6       3
3  bcc_83285|Cycle_2           180     105      41      15      26
4  bcc_83285|Cycle_2            81       2       5     100     150
5 bcc_161433|Cycle_2           127      20      17      60      58
6 bcc_161433|Cycle_2            80      58      33       7       1
```

To perform the differential analysis, a vector that describes the experimental plan is needed. In this case study, there are two replicates of the SK-N-SH cell line without treatment (SKNSH) followed by two replicates of the same cell line treated with retinoic acid (SKSNH-RA). So the `myConditions_AS` vector is defined as follows:

```
myConditions_AS <- c(rep("SKNSH", 2), rep("SKNSH-RA", 2))
```

4.1.3 Quality control

Before running the differential analysis, we check that the data was loaded correctly, using the `qualityControl` function.

```
qualityControl(myCounts_AS, myConditions_AS)
```

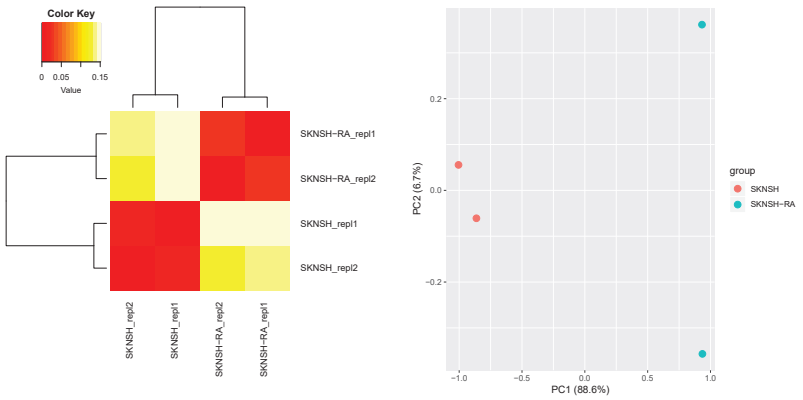

Figure 6: Quality control plots on alternative data
Left: Heatmap of the sample-to-sample distances for the alternative splicing dataset. Right: Principal Component Analysis for the alternative splicing dataset.

On both plots returned by the `qualityControl` function (Figure 6), the replicates of the same condition seem to be more similar between themselves than to the samples of the other condition. On the heatmap (left of Figure 6), the samples of the same condition cluster together. On the PCA plot (right of Figure 6), the first principal component (which summarises 88% of the total variance) clearly discriminates the two conditions.

4.1.4 Differential analysis

The main function of *kissDE*, `diffExpressedVariants`, can now be run to compute the differential analysis. Outputs are stored in a `myResult_AS` object (for a detailed description of its structure, see section 2.4.1) and the result for the first three events is given below:

```
myResult_AS <- diffExpressedVariants(myCounts_AS, myConditions_AS)
head(myResult_AS$finalTable, n = 3)
```

```
                                        ID Length_diff
bcc_83285|Cycle_2             bcc_83285|Cycle_2          99
bcc_52250|Cycle_0             bcc_52250|Cycle_0         160
bcc_135201|Cycle_433392 bcc_135201|Cycle_433392         104
                        Variant1_SKNSH_repl1_Norm Variant1_SKNSH_repl2_Norm
bcc_83285|Cycle_2                              84                         44
bcc_52250|Cycle_0                              10                         24
bcc_135201|Cycle_433392                        40                         29
                        Variant1_SKNSH-RA_repl1_Norm
bcc_83285|Cycle_2                                 17
bcc_52250|Cycle_0                                 15
bcc_135201|Cycle_433392                           19
```

```
                          Variant1_SKNSH-RA_repl2_Norm Variant2_SKNSH_repl1_Norm
bcc_83285|Cycle_2                                   28                          2
bcc_52250|Cycle_0                                   14                          2
bcc_135201|Cycle_433392                             27                          2
                          Variant2_SKNSH_repl2_Norm Variant2_SKNSH-RA_repl1_Norm
bcc_83285|Cycle_2                                 5                           110
bcc_52250|Cycle_0                                 0                            19
bcc_135201|Cycle_433392                           1                            32
                          Variant2_SKNSH-RA_repl2_Norm Adjusted_pvalue
bcc_83285|Cycle_2                                  162        0.00e+00
bcc_52250|Cycle_0                                   24        1.63e-06
bcc_135201|Cycle_433392                             59        1.88e-13
                          Deltaf/DeltaPSI lowcounts
bcc_83285|Cycle_2                  -0.809     FALSE
bcc_52250|Cycle_0                  -0.746     FALSE
bcc_135201|Cycle_433392            -0.715     FALSE
```

The first event in the `myResult_AS` output has a very low p-value (`Adjusted_pvalue` column, less than `2.2e-16`) and a very contrasted ΔPSI (`Deltaf/DeltaPSI` column, equal to `-0.804`) close to the maximum value (1 in absolute). This gene is differentially spliced. When the SK-N-SH cell line is treated with retinoic acid, the inclusion variant becomes the major isoform.

4.1.5 Export results

In order to facilitate the downstream analysis of the results, two tables are exported: the result table (`myResults_AS$finalTable` object, see section 2.4.1) is saved in a 'results_table.tab' file and the PSI table (`myResults_AS$f/psiTable`, see section 2.4.2) is saved in a 'psi_table.tab' file. Here are the commands to carry out this task:

```
writeOutputKissDE(myResults_AS, output = "results_table.tab")
writeOutputKissDE(myResults_AS, output = "psi_table.tab", writePSI = TRUE)
```

4.2 Application of *kissDE* to SNPs/SNVs

This second example present an analysis of SNPs/SNVs done with *kissDE* on RNA-Seq data from a subset of the case study presented in [2] (http://kissplice.prabi.fr/TWAS/).
The original purpose of this study was to demonstrate that the method can deal with pooled data (i.e. individuals are pooled prior to sequencing). Pooling can be used to decrease the costs. It is also sometimes the only option, when too few RNA is available per individual. The method can in principle be used on unpooled data, polyploid genomes, and for the detection of somatic mutations, but has for now only been evaluated for the detection of SNPs/SNVs in pooled RNAseq data.
In the remaining, we use the term SNV, which designates a variation of a single nucleotide, without any restriction on the frequency of the two alleles. The term SNP is indeed classically used for variants present in at least 1% of a population.

4.2.1 Dataset

The dataset comes from the human GEUVADIS project. Two populations were selected: Toscans (TSC) and Central Europeans (CEU). For each population, we selected 10 individuals, which are pooled in two groups of 5. Each group corresponds to a replicate for *kissDE*. The conditions being compared are the populations.

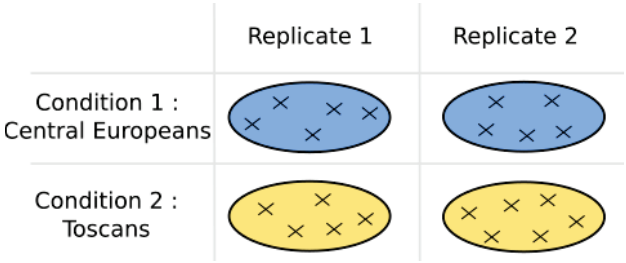

Figure 7: Experimental design of the SNP dataset
Each cross corresponds to an individual.

The data are paired-end. So each sample consists of 2 files. In total, 8 files have been used: 4 files for the two TSC samples and 4 files for the two CEU samples. Paired-end files from a same sample have been given as following each other to *KisSplice*.

KisSplice outputs a fasta file that stores SNVs found in the dataset. Its structure is described in section 2.1.3. The first SNV is presented below:

```
>bcc_44787|Cycle_421687|Type_0b|upper_path_length_131|C1_455|C2_455|C3_839|
C4_848|C5_5|C6_0|C7_39|C8_31|Q1_58|Q2_55|Q3_51|Q4_53|Q5_70|Q6_0|Q7_66|Q8_65|
rank_0.97008
CCAGAGAATCGGTCAGGGACCCCTGAGGGCCGCTGATTATTCCTATAGATGAGGAGTTTGGGGGCCGTTCCTGGGA
GCTGCTGGTACCAGTTTACAGTATTACTTCCGATGTTGGAGCTGCTTCCAGAACA
>bcc_44787|Cycle_421687|Type_0b|lower_path_length_131|C1_12|C2_14|C3_11|
C4_11|C5_18|C6_10|C7_4481|C8_4088|Q1_0|Q2_0|Q3_0|Q4_0|Q5_0|Q6_0|Q7_35|Q8_35|
rank_0.97008
CCAGAGAATCGGTCAGGGACCCCTGAGGGCCGCTGATTATTACTAGAGAAGAGGAGTTTGGGGGCCGTTCCTGGGA
GCTGCTGGTACCAATTTACAGTATTACTTCCGATGTTGGAGCTGCTTCCAGAACA
```

Events are reported in 4 lines, the two first represent one allele of the SNV, the two last the other allele. Thus the sequences only differ from each other at one position which corresponds to the SNV, here A/C in the center of the sequence (at position 42).

Because *KisSplice* was run with the default value of the `counts` parameter (i.e., 0), the counts have the following format `C1_x|C2_y|...|Cn_z`. In this example, there are 8 counts because we input 8 files. Each count corresponds to the reads coming from each file that could be mapped on the variant, in the order they have been passed to *KisSplice*. This information is particularly important in *kissDE* since it represents the counts used for the test.

4.2.2 Load data

The first step is to convert this fasta file (here called 'output_kissplice_SNV.fa') into a format that will be used in *kissDE* main functions, thanks to the `kissplice2counts` function.
Comment: `fileInSNV` contains the absolute path of the file on the user's hard disk.
Due to paired-end RNA-Seq data, the `pairedEnd` parameter was set to `TRUE`.

This conversion in a table of counts is stored in the `myCounts_SNV` object (for a detailed description of its structure, see section 2.1.3) and can be done as follows:

```
fileInSNV <- system.file("extdata", "output_kissplice_SNV.fa",
    package = "kissDE")
myCounts_SNV <- kissplice2counts(fileInSNV, pairedEnd = TRUE)
head(myCounts_SNV$countsEvents)
```

```
           events.names events.length counts1 counts2 counts3 counts4
1 bcc_44787|Cycle_421687           131     910    1687       5      70
2 bcc_44787|Cycle_421687           131      26      22      28    8569
3 bcc_44787|Cycle_421701           139     389    3349       2     149
4 bcc_44787|Cycle_421701           139      88      31      29    8821
5     bcc_100871|Cycle_3           107       0      10       0       0
6     bcc_100871|Cycle_3           107       3       1      13      10
```

To perform the differential analysis, a vector with the conditions has to be provided.
In the example, there are two replicates of TSC and two replicates of CEU, thus the condition vector `myConditions_SNV` is:

```
myConditions_SNV <- c(rep("TSC", 2), rep("CEU", 2))
```

4.2.3 Quality control

Before running the differential analysis, we recommand to check if the data was correctly loaded, by running the `qualityControl` function.

```
qualityControl(myCounts_SNV, myConditions_SNV)
```

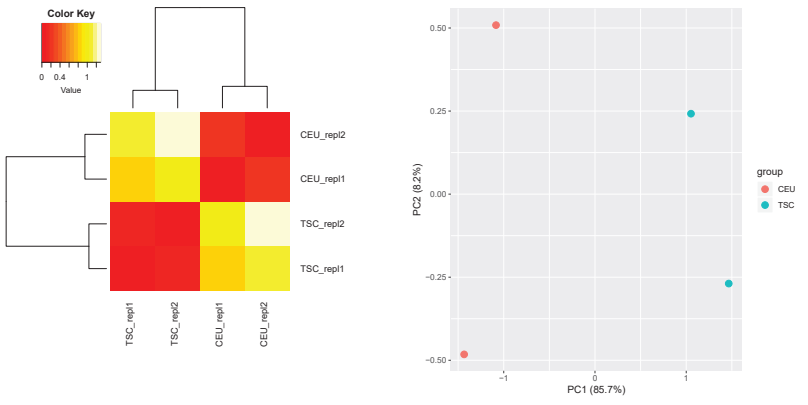

Figure 8: Quality control plots on SNV data
Left: Heatmap of the sample-to-sample distances on SNV data. Right: Principal Component Analysis on SNV data.

On both plots outputed (Figure 8), the replicates of the same condition seem to be more similar between themselves than to the samples of the other condition. On the heatmap (left of Figure 8), the samples of the same condition cluster together. On the PCA plot (right of Figure 8), the first principal component (which summarises 88% of the total variance) clearly discriminates the two conditions.

4.2.4 Differential analysis

The main function of *kissDE*, `diffExpressedVariants`, can now be run to compute the statistical test.
Outputs are stored in a `myResult_SNV` object (for a detailed description of its structure, see section 2.4.1) and the result for the first three events is printed:

```
myResult_SNV <- diffExpressedVariants(myCounts_SNV, myConditions_SNV)
head(myResult_SNV$finalTable, n = 3)
```

```
                                        ID Length_diff
bcc_44787|Cycle_320265 bcc_44787|Cycle_320265           0
bcc_100871|Cycle_3         bcc_100871|Cycle_3           0
bcc_44787|Cycle_421687 bcc_44787|Cycle_421687           0
                       Variant1_CEU_repl1_Norm Variant1_CEU_repl2_Norm
bcc_44787|Cycle_320265                    2014                    1172
bcc_100871|Cycle_3                           0                       0
bcc_44787|Cycle_421687                       5                      72
                       Variant1_TSC_repl1_Norm Variant1_TSC_repl2_Norm
bcc_44787|Cycle_320265                       0                       2
bcc_100871|Cycle_3                           0                      10
bcc_44787|Cycle_421687                     959                    1672
                       Variant2_CEU_repl1_Norm Variant2_CEU_repl2_Norm
bcc_44787|Cycle_320265                      23                     181
bcc_100871|Cycle_3                          12                      10
bcc_44787|Cycle_421687                      25                    8836
                       Variant2_TSC_repl1_Norm Variant2_TSC_repl2_Norm
bcc_44787|Cycle_320265                     179                     853
bcc_100871|Cycle_3                           3                       1
bcc_44787|Cycle_421687                      27                      22
                       Adjusted_pvalue Deltaf/DeltaPSI lowcounts
bcc_44787|Cycle_320265        0.00e+00          -0.926     FALSE
bcc_100871|Cycle_3            1.46e-04           0.909     FALSE
bcc_44787|Cycle_421687        1.85e-05           0.892     FALSE
```

The first event in the `myResult_SNV` output has a low p-value (`Adjusted_pvalue` column, equal to `8.63e-13`) and a very high absolute value of Δf (`Deltaf/DeltaPSI` column, equal to `-0.926`) close to the maximum value (1 in absolute). This SNP would typically be population specific. One allele is enriched in the Toscan population, the other in the European population.

4.2.5 Export results

We consider as significant the events that have an adjusted p-value lower than 5%, so we set `adjPvalMax = 0.05`. Results passing this threshold are saved in a '`final_table_significants.tab`' file, with the `writeOutputKissDE` function, as follows:

```
writeOutputKissDE(myResults_SNV, output = "final_table_significants.tab",
    adjPvalMax = 0.05)
```

4.3 Time / Requirements

The statistical analysis function (`diffExpressedVariants`) is the most time-consuming steps. Here is an example of the running time of this function on the two complete datasets presented in the case studies(section 4). The time presented were evaluated on a desktop computer with the following caracteristics: Intel Core i7, CPU 2,60 GHz, 16G RAM.

Dataset	Options	Number of events	Running time of `diffExpressedVariants`
AS data	`counts=2, pairedEnd=TRUE k2rg=TRUE`	59132	17m
SNV data	`counts=0, pairedEnd=TRUE`	64824	18m

Table 3: Profiling
Running time of the principal function of *kissDE* (`diffExpressedVariants`) for two datasets (AS dataset from the ENCODE project [10] described in section 4.1 and SNV dataset from the GEUVADIS project [12] described in section 4.2).

To reduce even more the running time of `diffExpressedVariants`, the parameter `nbCore` can be used to parallelize the most time-consuming step of this function (for more detailed explanation on this parameter see section 2.3).

5 Session info

```
sessionInfo()
```

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.2 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.10-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.10-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
```

```
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] kissDE_1.5.0

loaded via a namespace (and not attached):
  [1] bitops_1.0-6                matrixStats_0.54.0
  [3] DSS_2.33.0                  bit64_0.9-7
  [5] doParallel_1.0.14           bsseq_1.21.0
  [7] RColorBrewer_1.1-2          GenomeInfoDb_1.21.0
  [9] tools_3.6.0                 backports_1.1.4
 [11] R6_2.4.0                    KernSmooth_2.23-15
 [13] rpart_4.1-15                HDF5Array_1.13.0
 [15] Hmisc_4.2-0                 DBI_1.0.0
 [17] lazyeval_0.2.2              BiocGenerics_0.31.0
 [19] colorspace_1.4-1            permute_0.9-5
 [21] nnet_7.3-12                 tidyselect_0.2.5
 [23] gridExtra_2.3               DESeq2_1.25.0
 [25] bit_1.1-14                  compiler_3.6.0
 [27] Biobase_2.45.0              htmlTable_1.13.1
 [29] DelayedArray_0.11.0         labeling_0.3
 [31] rtracklayer_1.45.0          caTools_1.17.1.2
 [33] scales_1.0.0                checkmate_1.9.1
 [35] genefilter_1.67.0           stringr_1.4.0
 [37] digest_0.6.18               Rsamtools_2.1.0
 [39] foreign_0.8-71              R.utils_2.8.0
 [41] rmarkdown_1.12              aod_1.3.1
 [43] XVector_0.25.0              base64enc_0.1-3
 [45] pkgconfig_2.0.2             htmltools_0.3.6
 [47] limma_3.41.0                BSgenome_1.53.0
 [49] htmlwidgets_1.3             rlang_0.3.4
 [51] rstudioapi_0.10             RSQLite_2.1.1
 [53] DelayedMatrixStats_1.7.0    BiocParallel_1.19.0
 [55] gtools_3.8.1                R.oo_1.22.0
 [57] acepack_1.4.1               dplyr_0.8.0.1
 [59] RCurl_1.95-4.12             magrittr_1.5
 [61] GenomeInfoDbData_1.2.1      Formula_1.2-3
 [63] Matrix_1.2-17               Rcpp_1.0.1
 [65] munsell_0.5.0               S4Vectors_0.23.0
 [67] Rhdf5lib_1.7.0              R.methodsS3_1.7.1
 [69] stringi_1.4.3               yaml_2.2.0
 [71] SummarizedExperiment_1.15.0 zlibbioc_1.31.0
 [73] gplots_3.0.1.1              rhdf5_2.29.0
 [75] plyr_1.8.4                  grid_3.6.0
 [77] blob_1.1.1                  gdata_2.18.0
 [79] parallel_3.6.0              crayon_1.3.4
 [81] lattice_0.20-38             Biostrings_2.53.0
 [83] splines_3.6.0               annotate_1.63.0
 [85] locfit_1.5-9.1              knitr_1.22
 [87] pillar_1.3.1                GenomicRanges_1.37.0
 [89] codetools_0.2-16            geneplotter_1.63.0
```

```
 [91] stats4_3.6.0                XML_3.98-1.19
 [93] glue_1.3.1                  evaluate_0.13
 [95] latticeExtra_0.6-28         data.table_1.12.2
 [97] BiocManager_1.30.4          foreach_1.4.4
 [99] gtable_0.3.0                purrr_0.3.2
[101] assertthat_0.2.1            ggplot2_3.1.1
[103] xfun_0.6                    xtable_1.8-4
[105] survival_2.44-1.1           tibble_2.1.1
[107] iterators_1.0.10            GenomicAlignments_1.21.0
[109] AnnotationDbi_1.47.0        memoise_1.1.0
[111] IRanges_2.19.0              cluster_2.0.9
[113] BiocStyle_2.13.0
```

References

[1] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi:10.1186/s13059-014-0550-8.

[2] Hélène Lopez-Maestre, Lilia Brinza, Camille Marchet, Janice Kielbassa, Sylvère Bastien, Mathilde Boutigny, David Monnin, Adil El Filali, Claudia Marcia Carareto, Cristina Vieira, Franck Picard, Natacha Kremer, Fabrice Vavre, Marie-France Sagot, and Vincent Lacroix. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research*, 44(19):e148, 2016. doi:10.1093/nar/gkw655.

[3] Clara Benoit-Pilven, Camille Marchet, Emilie Chautard, Leandro Lima, Marie-Pierre Lambert, Gustavo Sacomoto, Amandine Rey, Cyril Bourgeois, Didier Auboeuf, and Vincent Lacroix. Annotation and differential analysis of alternative splicing using de novo assembly of rnaseq data. *bioRxiv*, 2016. URL: https://www.biorxiv.org/content/early/2016/09/12/074807, arXiv:https://www.biorxiv.org/content/early/2016/09/12/074807.full.pdf, doi:10.1101/074807.

[4] Gustavo A. T. Sacomoto, Janice Kielbassa, Rayan Chikhi, Raluca Uricaru, Pavlos Antoniou, Marie-France Sagot, Pierre Peterlongo, and Vincent Lacroix. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, 13(6):S5, 2012. doi:10.1186/1471-2105-13-S6-S5.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. URL: http://www.jstor.org/stable/2346101.

[6] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–43, 2013. doi:10.1093/biostatistics/kxs033.

[7] Hao Feng, Karen N. Conneely, and Hao Wu. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 42(8):e69, 2014. doi:10.1093/nar/gku154.

[8] Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin, and Karen N. Conneely. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, 43(21):e141, 2015. doi:10.1093/nar/gkv715.

[9] Yongseok Park and Hao Wu. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446, 2016. doi:10.1093/bioinformatics/btw026.

[10] Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Roder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigo, and Thomas R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012. doi:10.1038/nature11233.

[11] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. doi:10.1093/bioinformatics/bts635.

[12] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, 2013. doi:10.1038/nature12531.

III. Caractérisation du profil transcriptomique des patients TALS

Afin de dresser les profils d'expression et d'épissage des patients TALS, nous avons séquencé et réalisé le transcriptome de 9 échantillons biologiques provenant de 7 enfants ou fœtus TALS et 13 échantillons contrôles (enfant ou fœtus). Cette analyse a été réalisée pour chaque tissu à disposition : des fibroblastes (5 patients vs 8 contrôles), du liquide amniotique (3 patients vs 4 contrôles) et des cellules de lignées lymphoblastoïdes (LCL) (1 patient vs 1 contrôle, avec un replicat technique). L'ensemble de l'analyse s'est effectuée sur la version hg19 – GRCh37 de l'assemblage du génome humain (datant de 2014), et en utilisant les annotations de ensembl75 (version la plus récente pour ce génome de référence au moment du début du projet en 2014).

A. Résumé de l'analyse

Nous avons commencé par décrire, dans les cellules contrôles, les niveaux d'expression des gènes U12 et la qualité de l'épissage des introns U12 pour pouvoir ensuite les comparer aux gènes et introns U2. Nous avons ainsi montré que la majorité des gènes U12 est exprimée dans nos trois tissus et qu'ils ont un niveau d'expression semblable, que ce soit en comparant les gènes U12 d'un tissu à un autre, ou bien en comparant les gènes U12 aux gènes U2. Concernant l'épissage, nos données ne supportent pas l'hypothèse que les introns U12 soient moins bien épissés que les introns U2 adjacents, comme proposé dans une expérience sur des cellules humaines d'adultes (Niemela et al., 2014), et nous observons que l'épissage mineur parait moins efficace dans les LCL que dans les fibroblastes et amniocytes. Nous trouvons, dans l'ensemble des cellules contrôles, 11 cas d'épissage alternatif U12 (un intron U12 épissé à la place d'un autre intron U12) et, pour la première fois dans des cellules humaines et de manière physiologique, 21 exemples d'épissage alternatif de « U2/U12 twintrons » (Figure 25). Ces twintrons correspondent au chevauchement d'un intron U2 avec un intron U12, chaque forme alternative correspondant vraisemblablement à l'épissage du spliceosome majeur ou mineur.

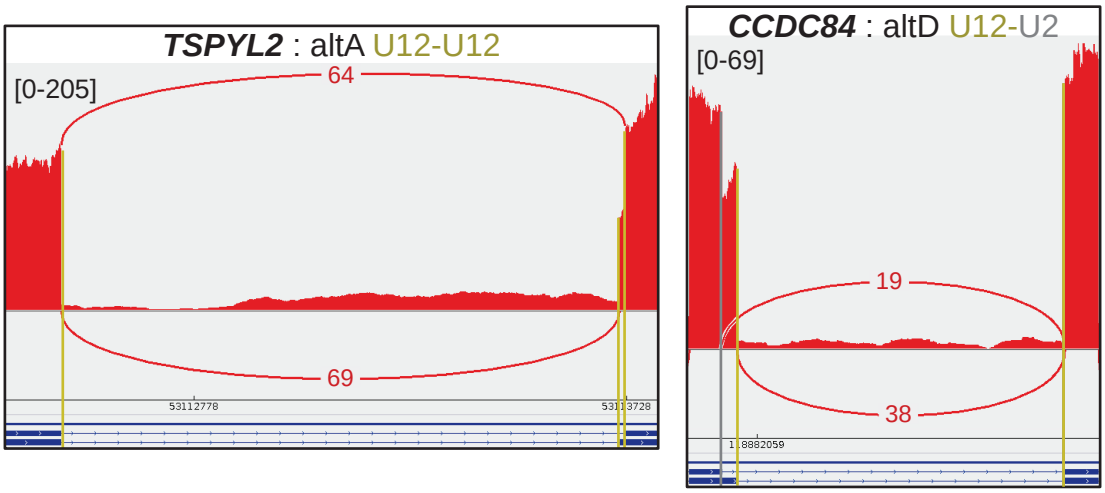

Figure 25 : Exemples d'événements d'épissages U12 physiologiques.
Sashimi plot des 14 échantillons contrôles. Les barres verticales jaunes et grises représentent des sites d'épissages U12 et U2, respectivement. Gauche : accepteur alternatif (événement U12-U12) dans le gène *TSPYL2*. Droite : donneur alternatif (événement sur un twintron U12-U2) dans le gène *CCDC84*. Les nombres indiquent la couverture moyenne par échantillon contrôle.

Concernant l'analyse patients vs contrôles, aucun différentiel d'expression des gènes U2 ou U12 n'a pu être mis en évidence. Cependant, de nombreuses rétentions d'introns U12-spécifiques sont détectées chez les patients, avec des effets très faibles dans les fibroblastes et amniocytes, et des effets forts dans les LCL. D'autres événements d'épissages sont observés, favorisant l'utilisation de sites U2 à proximité des sites U12 chez les patients, encore une fois de manière très prononcée chez les LCL où ces sites U2 sont souvent non-annotés.

Les gènes contenant les introns U12 les plus fortement affectés par ces événements d'épissage sont pour la plupart impliqués dans la division cellulaire, le développement des organes ainsi que l'immunité.

Ce travail a été publié en Juin 2019 dans la revue *RNA*. L'article et les figures supplémentaires sont inclus dans le paragraphe suivant. Les tableaux supplémentaires sont disponibles à l'adresse suivante : https://rnajournal.cshlp.org/content/early/2019/06/07/rna.071423.119/suppl/DC1.

B. Publication (*RNA*)

New insights into minor splicing—a transcriptomic analysis of cells derived from TALS patients

AUDRIC COLOGNE,[1,2] CLARA BENOIT-PILVEN,[1,2] ALICIA BESSON,[2] AUDREY PUTOUX,[2,3] AMANDINE CAMPAN-FOURNIER,[1,2] MICHAEL B. BOBER,[4] CHRISTINE E.M. DE DIE-SMULDERS,[5,6] AIMEE D.C. PAULUSSEN,[5,6] LUCILE PINSON,[7] ANNICK TOUTAIN,[8,9] CHAIM M. ROIFMAN,[10,11] ANNE-LOUISE LEUTENEGGER,[12,13] SYLVIE MAZOYER,[2,13] PATRICK EDERY,[2,3,13] and VINCENT LACROIX[1,13]

[1]INRIA Erable, CNRS LBBE UMR 5558, University Lyon 1, University of Lyon, F-69622 Villeurbanne, France
[2]"Genetics of Neurodevelopment" Team, Lyon Neuroscience Research Centre, UMR5292 CNRS U1028 Inserm, University of Lyon, F-69500 Bron, France
[3]Clinical Genetics Unit, Department of Genetics, Hospices Civils de Lyon, F-69500 Bron, France
[4]Division of Medical Genetics, Nemours/Alfred I. du Pont Hospital for Children, Wilmington, Delaware 19803, USA
[5]Department of Clinical Genetics, Maastricht University Medical Center, 6202 AZ Maastricht, The Netherlands
[6]School for Oncology and Developmental Biology, GROW, Maastricht University, 6229 ER Maastricht, The Netherlands
[7]Genetic Department for Rare Diseases and Personalized Medicine, Clinical Division, CHU Montpellier, F-34000 Montpellier, France
[8]Department of Genetics, Tours University Hospital, F-37000 Tours, France
[9]UMR 1253, iBrain, Tours University, Inserm, F-37000 Tours, France
[10]Department of Paediatrics, University of Toronto, Toronto, ON M5G 1X8, Canada
[11]Division for Immunology and Allergy, Canadian Center for Primary Immunodeficiency, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
[12]Université de Paris, NeuroDiderot, Inserm, F-75010 Paris, France

ABSTRACT

Minor intron splicing plays a central role in human embryonic development and survival. Indeed, biallelic mutations in *RNU4ATAC*, transcribed into the minor spliceosomal U4atac snRNA, are responsible for three rare autosomal recessive multimalformation disorders named Taybi–Linder (TALS/MOPD1), Roifman (RFMN), and Lowry–Wood (LWS) syndromes, which associate numerous overlapping signs of varying severity. Although RNA-seq experiments have been conducted on a few RFMN patient cells, none have been performed in TALS, and more generally no in-depth transcriptomic analysis of the ~700 human genes containing a minor (U12-type) intron had been published as yet. We thus sequenced RNA from cells derived from five skin, three amniotic fluid, and one blood biosamples obtained from seven unrelated TALS cases and from age- and sex-matched controls. This allowed us to describe for the first time the mRNA expression and splicing profile of genes containing U12-type introns, in the context of a functional minor spliceosome. Concerning *RNU4ATAC*-mutated patients, we show that as expected, they display distinct U12-type intron splicing profiles compared to controls, but that rather unexpectedly mRNA expression levels are mostly unchanged. Furthermore, although U12-type intron missplicing concerns most of the expressed U12 genes, the level of U12-type intron retention is surprisingly low in fibroblasts and amniocytes, and much more pronounced in blood cells. Interestingly, we found several occurrences of introns that can be spliced using either U2, U12, or a combination of both types of splice site consensus sequences, with a shift towards splicing using preferentially U2 sites in TALS patients' cells compared to controls.

Keywords: MOPD1; *RNU4ATAC*; minor splicing; U12-type introns; RNA sequencing; intron retention

INTRODUCTION

Pre-mRNA splicing is a crucial step that needs accurate execution for proper eukaryotic gene expression. Multiexonic pre-mRNA species can be spliced in a variety of ways as one or several exons may be skipped, introns retained or spliced with alternative donor or acceptor sites, either as part of a physiological process named alternative splicing or as the result of anomalies in the splicing process. Splicing misregulation may occur during cell

[13]These authors contributed equally to this work.
Corresponding author: sylvie.mazoyer@inserm.fr
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.071423.119. Freely available online through the *RNA* Open Access option.

proliferation, differentiation, survival or death, and is well documented in the context of numerous human diseases (Scotti and Swanson 2016).

Two types of introns coexist in the genome of most eukaryotes, major and minor introns (respectively also named U2- and U12-type introns) (Burge et al. 1998; Sheth et al. 2006). U12-type introns were first discovered due to their unusual AT-AC dinucleotide donor and acceptor splice sites and believed to harbor exclusively these sequences (Jackson 1991). They are now computationally identified based on their specific donor splice site and branch point sequence (BPS) consensus sequences, the latter being located within a specific window of 10–13 nt before the acceptor splice site (Dietrich et al. 2001a, 2005). In 2007, these criteria enabled to identify 695 introns of the U12-type in the human genome, thus representing <1% of all human introns (Alioto 2007). It turned out that 70% of these introns had the classical GT-AG termini.

Each type of intron is spliced by a distinct nuclear machinery: the major, or U2-dependent, spliceosome, and the minor, or U12-dependent, spliceosome. Both contain two small nuclear ribonucleoproteins (snRNPs) involved in intron recognition (respectively, U1 and U2, and U11/U12 di-snRNPs) and three snRNPs involved in the catalytic reaction (respectively, U4/U6.U5 and U4atac/U6atac. U5 tri-snRNPs), U5 being the only snRNA shared between the two spliceosomes. While major spliceosome- and minor spliceosome-specific snRNAs have divergent sequences, they share a similar secondary structure (Tarn and Steitz 1996). Spliceosome specificity relies mostly on splice sites recognition by the major U1 and U2 snRNPs or the minor U11/U12 di-snRNP, and protein composition of the two spliceosomes is highly similar apart from seven proteins which are specific to the minor spliceosome (Schneider et al. 2002).

Minor splicing conservation through evolution implies an important role for this cellular process, but a more direct evidence of its central role came with the identification of mutations in a component of the minor spliceosome in patients afflicted with a severe developmental disease. Indeed, an autosomal recessive disorder named microcephalic osteodysplastic primordial dwarfism type 1 (MOPD1, OMIM 210710) or Taybi–Linder syndrome (TALS) was found by our team and others to be due to biallelic mutations in the gene transcribed into U4atac, *RNU4ATAC* (Edery et al. 2011; He et al. 2011). This very rare syndrome is characterized by multiple malformations including severe microcephaly and cortical brain malformations (neuronal migration defects), corpus callosum agenesis/dysgenesis, cerebellar vermis hypoplasia, intellectual disability, dysmorphic features, sparse or absent hair, dry skin, short stature and bone anomalies. It leads to early unexplained death occurring within the first three years of life in more than 70% of the cases. Interestingly, other very rare congenital disorders, namely Roifman syndrome (RFMN, OMIM 616651) and Lowry–Wood syndrome (LWS, OMIM 226960) have also been recently attributed to biallelic *RNU4ATAC* mutations (Merico et al. 2015; Farach et al. 2018). Both RFMN and LWS have features overlapping with TALS (i.e., microcephaly, intellectual deficiency, growth retardation, skeletal dysplasia) but these disorders are not associated with early mortality, they do not include visible structural brain anomalies, and they have less pronounced microcephaly and growth retardation. Of note, RFMN cases exhibit a specific antibody deficiency that is the hallmark of this rare immunodeficiency syndrome.

The U4atac/U6atac bi-molecule has a Y-shaped structure which consists of two intermolecular stems, stem I and stem II, separated by a secondary U4atac structure called the 5′ stem–loop. The U4atac terminal region also contains a 3′ stem–loop and a Sm protein-binding site (for review, see Turunen et al. 2013). To date, mutations have been identified at the homozygous or compound heterozygous states in *RNU4ATAC* in 53 TALS, 14 RFMN and 5 LWS patients or fetuses (from 30 TALS, 10 RFMN, and 4 LWS families, respectively) (Ferrell et al. 2016; Putoux et al. 2016; Bogaert et al. 2017; Dinur Schejter et al. 2017; Farach et al. 2018; Hallermayr et al. 2018; Heremans et al. 2018; Lionel et al. 2018; Shelihan et al. 2018; Wang et al. 2018; Shaheen et al. 2019). Quite clear, although preliminary, phenotype–genotype correlations stand out across the growing number of cases: Early death in TALS patients (usually before 3 yr of age) is associated with homozygosity for the most common pathogenic variant, g.51G > A, located in the 5′ stem–loop which contains most of the TALS mutations; RFMN is always associated with the location of at least one of the two mutations in Stem II, a region never found mutated in TALS patients.

While germline mutations in genes encoding core protein components of the spliceosome had been already involved in genetic diseases (some forms of retinitis pigmentosa and rare craniofacial, skeletal and skin disorders), U4atac was the first spliceosomal snRNA in which mutations were identified (for reviews, see Padgett 2012; Verma et al. 2018). Since then, mutations in *RNU12* were associated with early onset cerebellar ataxia in a large consanguineous family (Elsaid et al. 2017). Mutations in spliceosome components are expected to cause global splicing dysregulation that should manifest in most, if not all tissues, an assumption difficult to reconcile with the highly restricted phenotypes observed in spliceosomopathies. Despite recent technological advances allowing in-depth analyses at the transcriptomic level, very few RNA-seq studies have been performed in these pathologies, precluding comprehensive description of the molecular events associated with the identified mutations. There is now a total of three published analyses of RNA-seq data from RFMN patients that revealed massive U12-type intron retention (IR), but each study focused on only two patients and was restricted to a single cell type, either mononuclear

blood cells or megakaryocytes (Merico et al. 2015; Dinur Schejter et al. 2017; Heremans et al. 2018). In contrast, the transcriptomic profile of TALS patients has not been described yet.

We present here for the first time the analysis of RNA-seq data sets performed on cells derived from skin biopsies, amniotic fluids and peripheral blood taken from seven unrelated TALS cases carrying various *RNU4ATAC* mutations and 13 control individuals matched for tissue, age and gender, hence providing the first whole genome splicing pattern and expression data for this disease. The thorough analysis of this unique data set enables us to study how minor splicing is carried out in physiological and pathological conditions, in various cell types, and sheds new light on this cellular process.

RESULTS

Presentation of RNA-seq data generation and analysis

Biological samples

A total of nine biological samples, that is, five skin, three amniotic fluid and one peripheral blood biospecimens, were obtained from seven unrelated previously published TALS cases (Table 1). This represents the largest collection of TALS samples, to the best of our knowledge. Among these seven cases, four (three children, one fetus) are homozygous for the most common *RNU4ATAC* mutation, g.51G > A, and three (one child, two fetuses) are compound heterozygous for g.50G > C;g.51G > A, g.40C > T; g.124G > A, and g.51G > A;g.124G > A, respectively. All the affected children died before the age of three, regardless of their mutation(s). Importantly, two different biospecimens were obtained for two g.51G > A homozygous patients, skin and blood for one child and maternal amniotic fluid and skin for the other. Biological samples (eight skin, four amniotic fluid, and one peripheral blood samples) were also obtained from 13 age- and sex-matched controls (Table 1).

RNA-seq protocol

We extracted total RNA from fibroblasts (derived from the skin biopsies), amniocytes (derived from the amniotic fluids), and lymphoblastoid cell lines (LCL, established by EBV immortalization of B lymphocytes obtained from blood samples). RNA-seq data were then generated in two experimental setups by Illumina sequencing of (1) poly(A)-selected, non strand-specific sequencing libraries (100 nt paired-end reads) on three patient samples in the pilot study; (2) poly(A)-selected, strand-specific sequencing libraries (75 nt paired-end reads) in the extended study. The extended study was technically more comprehensive and comprised all the samples which had been sequenced in the pilot study. Consequently, we will present and discuss the results obtained in this latter study only. However, in the LCL in-depth analysis, we also used the data set of our pilot experiment as a technical replicate, in order to make up for the lack of biological replicates.

Data sets analysis

Our analysis of these 24 transcriptomes (nine patient and 13 control data sets from the extended study; one patient and one control LCL data sets from the pilot study) examined both gene expression and splicing alterations with a special focus on IR. To this aim, we set up three bioinformatic pipelines (see Materials and Methods), that is, a bioinformatic pipeline that uses a recently developed mapping-first approach dedicated to accurate IR detection, IRFinder (Middleton et al. 2017), and two other pipelines that allow us to identify other types of alternative splicing events, one with a mapping-first approach, vast-tools (Tapial et al. 2017), and the other with an assembly-first approach that we previously reported to have the ability to detect the use of unannotated splice sites, KisSplice (Benoit-Pilven et al. 2018a). Statistical significance of the results obtained with these three pipelines was determined using the same analytical tool, kissDE (Benoit-Pilven et al. 2018b), which allows the identification of significant changes in relative intron or exon inclusion across conditions. To quantify the magnitude of the changes, we computed the Percent Spliced In (PSI) metric, which is the ratio of the reads including the intron over the sum of the reads including or excluding it, for each intron and each condition. This metric provides values close to 100% for fully retained introns and to 0% for fully spliced introns. The PSI metric was also used for quantifying other types of alternative splicing events (see Materials and Methods). The difference between conditions, $\Delta PSI = PSI\_Patients - PSI\_Controls$, is a measure of the magnitude of the splicing alteration; the sign of this metric indicates in which condition the retention is more frequently seen (patients for positive values or controls for negative values), and its absolute value indicates the level of the difference (the closer to 100%, the higher the difference).

All U12-type intron alternative splicing events identified in patients' cells are reported in Supplemental Table S1 and described in details in Supplemental Table S2. The processed underlying data can be explored in a Shiny Interface at http://lbbe-shiny.univ-lyon1.fr/TALS-RNAseq/.

Expression levels and splicing efficiency of U12 genes in control fibroblasts, amniocytes and LCL

Global mRNA expression levels of U12 genes in control cell types

To date, despite the large number of transcriptomic studies performed in human tissues and cell types, the spatial

TABLE 1. Description of the samples analyzed by RNA-seq

	Biological sample	Analyzed cells	RNA-seq experiment(s)	*RNU4ATAC* pathogenic variants	Age at sample collection	Age at death	Gender	Patient identification
TALS collection	Skin biopsy	Fibroblasts	Pilot study + Extended study	g.51G > A ; g.51G > A	2 mo	28 mo	F	TALS6 (Edery et al. 2011)
				g.51G > A ; g.51G > A	10 mo (post-mortem)	10 mo	M	TALS2 (Edery et al. 2011)
				-	2 mo	-	F	-
				-	21 mo	-	M	-
			Extended study	g.51G > A ; g.51G > A	4 mo	7 mo	F	TALS4 (Edery et al. 2011)
				g.50G > C ; g.51G > A	29 mo	29 mo	F	TALS10 (Edery et al. 2011)
				g.40C > T ; g.124G > A	30 GW (post-mortem)	30 GW (TOP)	M	Fetus 3 (Putoux et al. 2016)
				-	7 mo	-	F	-
				-	39 mo	-	F	-
				-	3 yr	-	F	-
				-	26 GW	-	M	-
				-	12 mo	-	M	-
				-	12 d	-	M	-
	Amniotic fluid	Amniocytes	Extended study	g.51G > A ; g.124G > A	21 GW	21 GW (TOP)	F	Fetus 2 (Putoux et al. 2016)
				g.51G > A ; g.51G > A	25 GW	25 GW (TOP)	F	Fetus 1 (Putoux et al. 2016)
				g.51G > A ; g.51G > A	20 GW	10 mo	M	TALS2 (Edery et al. 2011)
				-	21 GW	-	F	-
				-	25 GW	-	F	-
				-	22 GW	-	M	-
				-	26 GW	-	M	-
	Peripheral blood	LCL	Pilot study + Extended study	g.51G > A ; g.51G > A	2 mo	28 mo	F	TALS6 (Edery et al. 2011)
				-	2 mo	-	F	-
RFMN collection	Peripheral blood	MBC	Merico et al. 2015	g.13C > T ; g.37G > A	38 yr	-	M	k1.p2 (Merico et al. 2015)
				g.13C > T ; g.48G > A	21 yr	-	M	k2.p3 (Merico et al. 2015)
				g.13C > T ; -	43 yr	-	M	-
				g.13C > T ; -	67 yr	-	M	-
				g.13C > T ; -	57 yr	-	M	-

M, male; F, female; GW, gestational weeks; TOP, termination of pregnancy; LCL, lymphoblastoid cell line; MBC, mononuclear blood cells.

and temporal pattern of expression of the transcribed genes containing at least one U12-type intron (hereafter called U12 genes, while U2 genes are those not containing any U12-type intron) has never been described and is largely unknown. We therefore first evaluated which U12 genes were expressed in the eight fibroblast, four amniocyte and one LCL samples derived from control children and fetuses to set the frame of reference for the comparison with TALS patients. We based our analysis on the set of 699 genes containing at least one U12-type intron that we identified in the human genome through a computational scan of the latest annotation of the GRCh37 assembly (Ensembl Release 75) with a U12-type intron annotation tool (Alioto 2007) (846 minor introns annotated in total, Supplemental Table S3), and fixed a threshold for expression at a mean of 5 Transcripts Per Million (TPM) both for U2 and U12 genes. Among these 699 genes, 528 (76%) are expressed in at least one cell type in our control data sets and 427 (61%) are expressed in the three of them, suggesting that the majority of the U12 genes are expressed in various cell types (Supplemental Fig. S1A). The distribution of the expression levels of U12 genes is highly similar between the three cell types and shows a peak at around 30 TPM (Supplemental Fig. S1B). However, we found that the mean number of transcripts per U12 gene was higher in the LCL than in amniocytes and fibroblasts (56, 51 and 48 TPM, respectively). When considering U2 genes, an extra peak of genes expressed

at a level <1 TPM is seen (Supplemental Fig. S1B); this bimodal distribution has already been reported and most likely corresponds to noise from the transcriptional machinery (Hebenstreit et al. 2011). Principal component analysis (PCA) of the expression levels of U12 and U2 genes demonstrated that the control transcriptome data sets partitioned depending on the cell type (Supplemental Fig. S1C), indicating that the U12 genes expression level pattern is specific to each cell type. Gender or prenatal vs. post-natal origin of the skin biopsies from which fibroblasts were derived did not strongly influence U2 and U12 genes expression level patterns (Supplemental Fig. S1C).

Global U12- and U2-type intron retentions in U12 genes

After examining mRNA expression levels, we focused on introns and their splicing efficiency in fibroblasts, amniocytes and LCL by analyzing the extent of IR using PSI value calculations. To alleviate potential biases due to the large difference in the number of U2 and U12 genes, we chose to restrict the analysis comparing U12- and U2-type intron splicing efficiency to introns located in U12 genes. In order to obtain robust PSI estimations, we focused on intronic regions with sufficient read coverage (i.e., number of exon-intron + exon–exon junction reads ≥10 in at least 4/8 fibroblast, 2/4 amniocyte, and the LCL control samples). The few annotated introns that were never found spliced out in our data sets were also removed (see Materials and Methods). The analysis was performed on a set of 366 U12-type introns and 1887 U2-type introns scattered in 337 U12 genes with a mean expression of at least 5 TPM in each cell type. We found that the mean PSI for the U12-type introns is 2.2% (median = 0.7%) in fibroblasts, 2.7% (median = 0.9%) in amniocytes and 4.4% (median = 2.0%) in LCL, whereas the mean PSI for the U2-type introns are respectively 3.9% (median = 1.1%), 4.7% (median = 1.5%), and 4.8% (median = 1.5%) in these cells. In contrast with a previous result obtained with HEp-2 cells (Niemela et al. 2014), we did not observe in our data sets that U12-type introns were spliced less efficiently than their neighboring U2-type counterparts. We further observed that splicing was most efficient in fibroblasts, and that U12-type intron splicing was less efficient in LCL (Supplemental Fig. S2A). PCA of the PSI values for U12- and U2-type introns also separated cell types, although less clearly than expression values as one of the four amniocyte data sets segregated with fibroblasts consistently in both U12- and U2-type introns analyses (Supplemental Fig. S2B). We noticed that the LCL data set singled out in PCA of U12-type IR, as it does in PCA of U12 gene expression levels, a finding confirmed when incorporating the pilot study data set in the analyses.

U12-type intron alternative splicing

Besides IR, more complex patterns of U12-type intron alternative splicing have on some occasions also been observed, although less frequently than for U2-type introns (Levine and Durbin 2001; Chang et al. 2007). To identify these events in our data sets we used both a mapping-first approach (vast-tools) and an assembly-first approach (KisSplice), as we previously showed that these approaches were complementary (Benoit-Pilven et al. 2018a). We focused on events with sufficient read coverage (same filter as that used for IR) and with exon–exon junctions covered by an average of at least five reads. We found 9 U12 genes for which a total of 10 complex minor splicing patterns were observed through the use of alternative U12 splice sites in all control data sets. In 9/10 cases, an alternative U12 acceptor site was used, leading to exon skipping in a few instances, while in the remaining one, both alternative U12 donor and acceptor sites were used. The use of the least common donor and/or acceptor splice sites was supported by more than 10% of the reads in all three cell types for six of these events, indicating that they are not marginal. It should be noted that half of the splicing events produced alternative forms considered as noncoding in databases because they contain premature termination codons (PTCs).

U12/U2 splice site switching

Most interestingly, we also found U12-type introns for which nearby U2 splice site(s) were sometimes favored over U12 splice site(s), probably in the context, in most cases, of a switch from the minor to the major spliceosome for splicing the intron. This phenomenon was first described for the *D. melanogaster prospero* gene (Scamborova et al. 2004); lately, the existence of these introns called U2/U12-type twintrons was extended to several other U12 genes in different species, including humans (for review, see Hafez and Hausner 2015). We identified 21 of such alternative events comprising or not the skipping of an exon in 16 U12 genes. In four of these events, both U2 alternative donor and acceptor splice sites were used. In 10 of them, a U2 alternative donor site was used in combination with the U12 acceptor site, and in the remaining seven, a U12 donor was used with an alternative U2 acceptor site. Such mixed patterns had not yet been observed, to the best of our knowledge. In 13/21 cases, the least abundant form represented more than 10% of all the reads in all three control cell types. A striking example of this situation was observed for the *CCDC84* gene (Supplemental Fig. S2C), for which the transcripts derived from the use of U12 splice sites (producing PTC-containing transcripts) or U2 donor and U12 acceptor splice sites (coding the full length protein of unknown function) are found in similar abundance. Hence, the type of splice sites selected to remove this intron from the *CCDC84* pre-mRNA can regulate

the amount of the full length protein which is produced without changing the transcriptional expression level of the gene.

Overall, beyond the description of these novel mechanisms, this first global analysis of U12-type intron splicing in cells from control children and fetuses provides a reference for studying the consequences of *RNU4ATAC* biallelic mutations on the transcriptome of cells derived from TALS patients.

Global impact of *RNU4ATAC* biallelic mutations on transcriptomes of fibroblasts, amniocytes and LCL derived from TALS patients

U12 mRNA expression levels in patients and controls

The PCA performed on either U2 or U12 genes expression levels (TPM measures) in the 22 data sets of the extended RNA-seq study (nine patient and 13 control samples) separated cell types again but failed to separate patients from controls (Supplemental Fig. S3). The fact that we did not see any global impact of *RNU4ATAC* mutations on U12 gene expression levels using PCA was surprising because one could expect that IR would trigger transcript degradation through quality control pathways, which would in turn lower their amount. These quality control pathways dealing with transcripts with retained introns could be the Nonsense Mediated mRNA Decay (NMD) acting in the cytoplasm (Wong et al. 2013, 2016), or could include both exosome-mediated mRNA turnover following nuclear sequestration, and NMD (Braunschweig et al. 2014). More specifically, U12-type IR have been shown to lead to nuclear retention and nuclear decay by the RNA exosome (Niemela et al. 2014). In order to investigate U12 genes expression levels further, we ran DESeq2 on the fibroblast data sets from controls and patients to identify differentially expressed (DE) genes (i.e., genes for which the number of produced polyA+ mRNAs differs). Using standard cutoffs, that is, False Discovery Rate (FDR) $\leq 5\%$ and $|\log_2(FC)| \geq 1$, we found only 13 DE genes (eight up-, five down-regulated), none of them containing any U12-type intron. The same analysis performed in the patient amniocyte data set collection produced a list of 32 DE genes (11 up-, 21 down-regulated), again all U2 genes, and all different from those identified in fibroblasts except one [*RP11-305K5*, $\log_2(FC) = 1.6$]. To evaluate the biological relevance of these DE genes, we calculated how many were identified in our RNA-seq fibroblast data sets in every possible combination of patients and age- and sex-matched controls and found that this number markedly decreased with the increasing number of patient samples (Supplemental Fig. S4). This pattern is similar to that obtained with false negative results in a study evaluating the number of biological replicates needed to ensure detection of valid significantly differentially expressed genes (Schurch et al. 2016). We therefore conclude that the DE genes we identified here are likely not associated to the pathology itself.

U12-type intron splicing in patients and controls

When performing PCA on U12-type IR levels using PSI values, we observed a clear partitioning of the patients and the controls, as expected, while the same analysis on U2-type IR failed to separate patients from controls (Fig. 1, top and bottom left). Axis 1 of the U12-type IR PCA (PC1: 88% of the variance) was essentially supported by LCL, showing that this cell type has a specific "sensitivity" to defects in U12-type intron splicing. Nevertheless, even when removing LCL data from the analysis, we find that the partition between patients and controls remains clear (Fig. 1, bottom right). We can thus conclude from the PCA analyses that U12-type intron splicing appears indeed globally altered in TALS patients, and that the splicing default appears somehow different in the LCL compared to fibroblasts and amniocytes. We next looked into more details at the global splicing anomalies associated with *RNU4ATAC* mutations in each cell type.

Splicing efficiency of U12- and U2-type introns in fibroblasts and amniocytes derived from TALS patients

Because the separate analysis of TALS fibroblasts and amniocytes produced similar results, we present them together. The fibroblast data sets (F) were obtained from three homozygous g.51G > A patients, two compound heterozygous g.51G > A;g.50G > C and g.40C > T;g.124G > A patients and eight controls; the amniocyte data sets (A) from two homozygous g.51G > A patients, one compound heterozygous g.51G > A;g.124G > A patient and four controls.

U2-type intron retentions

As expected, the mean PSI values for the U2-type introns passing our filters (see Materials and Methods) were similar in patients and controls [respectively 4.6% vs. 4.3% (F) and 5.1% vs. 5.2% (A)], suggesting that the TALS patients' cells exhibit unchanged U2-type intron splicing profiles. Indeed, a very small fraction of U2-type introns were found markedly retained ($\Delta PSI \geq 10\%$ and $FDR \leq 5\%$) in patients: 79 out of 54922 (F); 133 out of 59255 (A). Only eight of them were found in both data sets, six of which occurring in U12 genes. As the current annotation is conservative and splice sites that show poor homology with U2- and U12-type intron consensus sequences tend to be considered U2-type, we suspected that some of the retained "U2-type introns" could be misclassified and should be reclassified as U12-type introns. Indeed when examining them, we identified four introns with non consensus splice site sequences located within the *RECQL5*, *DERL2*,

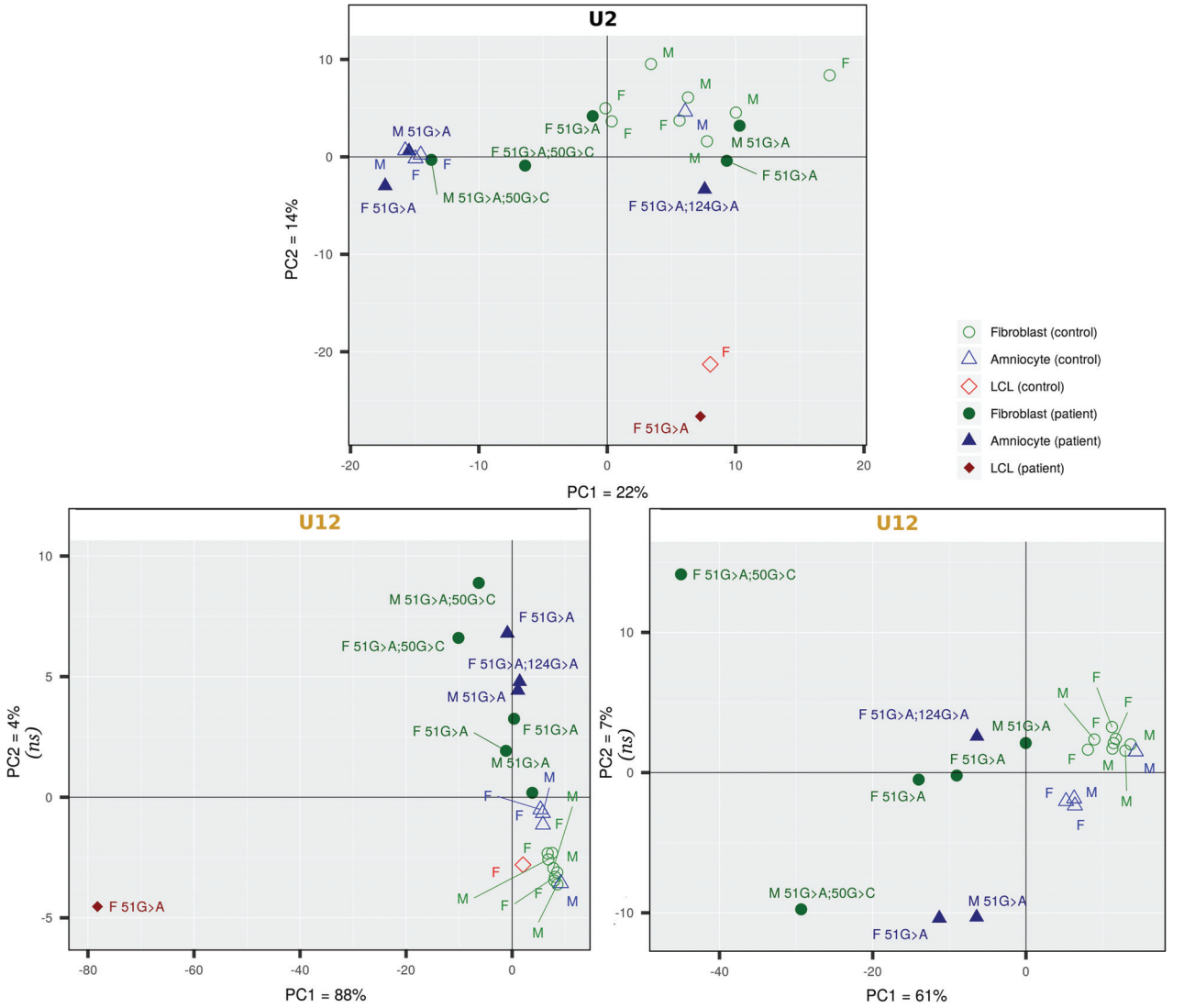

FIGURE 1. Patterns of U2- and U12-type IRs in TALS patient and control cells. Principal component analyses of the most variable mean PSI values of U2- and U12-type introns are presented. PCA for U12-type introns was performed with (*left*) and without the LCL data sets (*right*). Fibroblasts, amniocytes and LCL were derived from tissues taken from control or TALS fetuses and children. The sex of the donor from which was derived each sample is indicated (M, Male; F, Female), as well as the *RNU4ATAC* mutation(s) for the patients' samples. (*ns*) not significant (the percentage of variance explained by the axis is smaller or equal to the percentage of variance expected by chance, see Materials and Methods).

KIAA0556 and *LZTR1* genes (Fig. 2 left, red dots; Supplemental Table S3; Supplemental Fig. S5). For these atypical donor or acceptor splice site sequences, at least one score regarding both U12- and U2- type introns is inferior to −1 (scores are log-likelihood ratios: A sequence with a negative score resembles more the background sequence than the consensus one). Such borderline cases are difficult to classify, and depending on the genome annotation used, their score slightly increases or decreases, causing the classifier to call them U12 or U2. Of note, they were originally classified as U12- or U2/U12-types in U12db (which uses U12 classification scripts, genomic sequences and annotations of 2007), not only in humans but also in macaques and chimpanzees for *RECQL5* and in many other species including zebrafish and mouse for *DERL2*, *LZTR1*, and *KIAA0556*. The fact that these introns are markedly retained in TALS samples strongly suggests that they are genuine U12-type introns, consistent with their higher scores for U12 compared to U2 splice site sequences. We thus propose to reclassify them as U12-type introns.

U12-type intron retentions

PSI computation was achieved for 482 (F) and 430 (A) U12-type introns, including the four introns that we previously

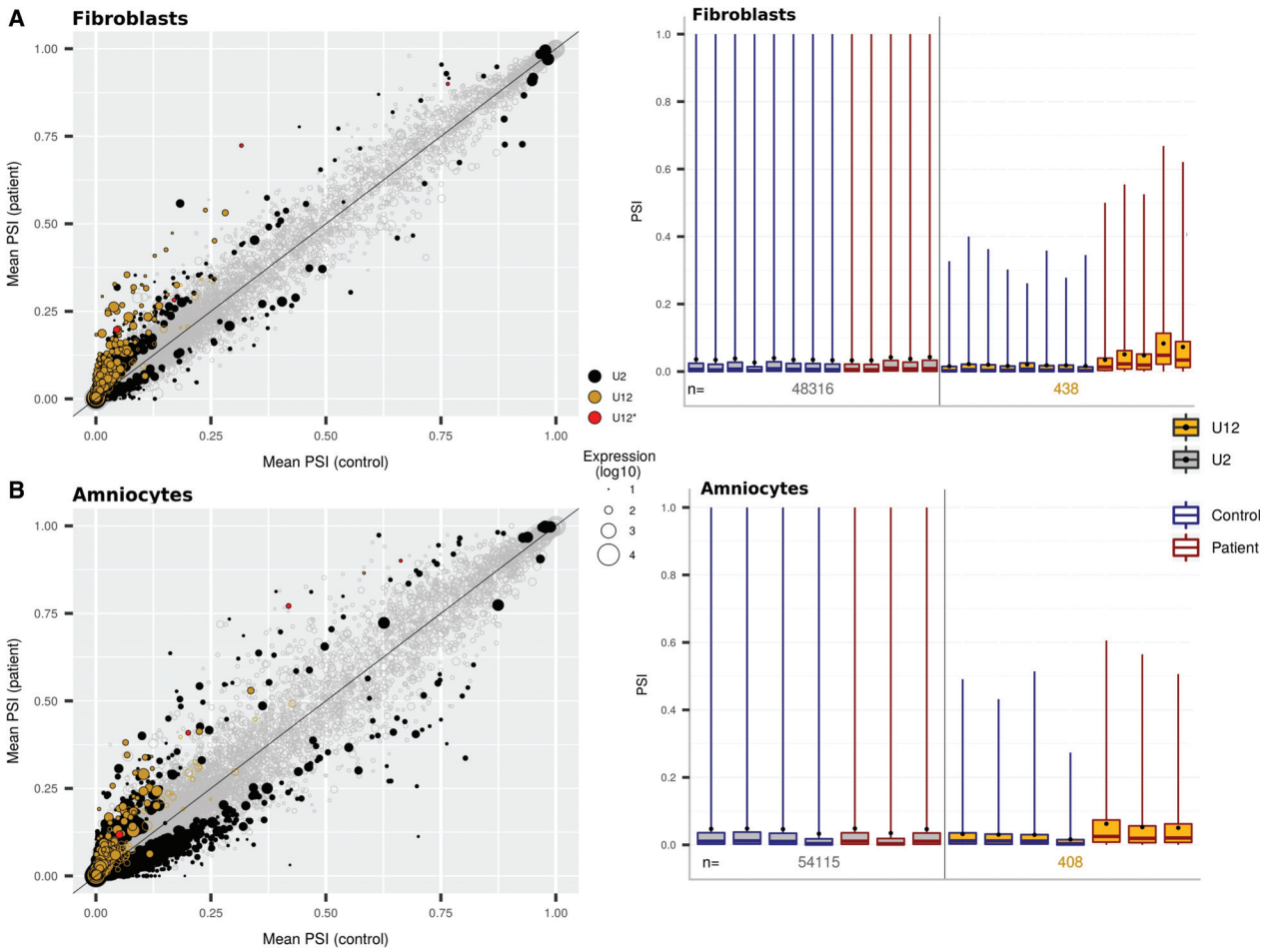

FIGURE 2. Comparison of U2- and U12-type IR levels in TALS patient and control cells. Analysis of the (*A*) fibroblast data sets or (*B*) amniocyte data sets. (*Left* panels) Plots of the mean U2- and U12-type IR levels expressed with the Percent Spliced In (PSI) metric and obtained for the patients' versus the controls' data sets (PSI-plots). Each circle represents an intron: the color indicates its type (U12* means U2-type intron proposed to be reclassified as U12-type in this study), the size indicates the amount of the corresponding transcript, and the filing status indicates the significance of the IR level (filled circle: FDR ≤ 5%; unfilled circle: FDR > 5%). The intron position respective to the line indicates whether the intron is more retained in patients (*above* the line) or controls (*below* the line). The further a point is from this line, the greater the intron's ΔPSI. (*Right* panels) Boxplots of U2- and U12-types intron PSI values of each patient's and control's data set (PSI-boxplots). Mean values are represented as black dots. The numbers of U2- and U12-type introns indicated correspond to those with robust PSI estimation and sufficient coverage in each sample.

reclassified as U12. As expected, we found that the mean PSI was higher for patients (~6%) than for controls (~3%) in both fibroblasts and amniocytes (Table 2; Fig. 2, right), testifying that minor splicing was indeed impaired in patients. Of note, mean PSI were higher in fibroblasts derived from the two *RNU4ATAC* compound heterozygous patients (Fig. 2A right, last two boxplots; Supplemental Table S4) than in the homozygous patients. However, surprisingly, the magnitude of the effect was limited. As a matter of fact, the vast majority of U12-type introns were statistically significantly misspliced (FDR < 5%), but most of them were only marginally affected (ΔPSI < 10%) (Table 2). The larger fraction of statistically significant U12-type IR observed in fibroblasts compared to amniocytes is most likely due to the larger patient/control sample set in the former (13 vs. 7, respectively), hence increasing the statistical power and enabling us to find more statistically significant small effects.

Concomitant U12- and U2-type intron retentions

For some U12-type IR, we noticed that the 5′ or 3′ neighboring U2-type intron was also retained. The example of the *DYNC1LI2* gene is given in Figure 3, top. The analysis of the 55 (F) and 33 (A) U12-type marked IR revealed respectively nine and three instances of concomitant U12- and neighboring U2-type IRs suggesting that the missplicing of some U12-type introns could lead to the

TABLE 2. Summary of the U12-type introns results from TALS and RFMN patients' data sets compared to controls' data sets

Data sets	TALS fibroblasts	TALS amniocytes	TALS LCL	RFMN MBC
Number of patients versus controls	5 versus 8	3 versus 4	1 × 2 versus 1 × 2	2 versus 3
Number of tested U12-type introns	482	430	480	285
Mean PSI patients versus mean PSI controls	6.7% versus 2.4%	6.4% versus 3.3%	27.5% versus 4.8%	28.7% versus 6.0%
Number of not retained or not significantly retained (FDR > 5%) U12-type introns	100	242	12	17
Number of significantly retained (FDR ≤ 5%) U12-type introns	382	188	468	268
Number of marked (|ΔPSI| ≥ 10%) and significantly retained (FDR ≤ 5%) U12-type introns	55	33	370	208
Mean ΔPSI	17.8%	17.5%	27.6%	28.9%

x2, technical replicates.

retention of the 5′ or 3′ adjacent U2-type intron. The scores of the splice sites of these U2-type introns are not different from those of the other U2-type introns, and they have weak U12 splice site scores (Supplemental Fig. S5, black points). Interactions between the minor and the major spliceosomes have already been suggested (Wu and Krainer 1996; Lewandowska et al. 2004; Tapial et al. 2017; Horiuchi et al. 2018), and our study provides additional observations further supporting this hypothesis.

Fibroblasts and amniocytes obtained from the same patient

We took advantage of the availability of both amniocytes and post-natal fibroblasts for a homozygous g.51G > A patient to assess U12- and U2-type IR in the same genetic background: This analysis revealed that the mean PSI and the PSI distributions of both U2- and U12-type introns were similar in these two cell-types (Supplemental Fig. S6, left).

Other types of U12-type intron alternative splicing

Vast-tools and KisSplice identified respectively four and two U12/U2 splice site switchings in the fibroblast and amniocyte data sets, all of them in favor of the use of U2 splice sites in TALS patients. In particular, both cell types exhibited the same splice site switching event in the *CCDC84* gene (shown for the fibroblast data sets in Fig. 3, bottom). The balance of the transcripts derived from the use of either the U12- or the U2-type splice sites observed in controls (~65%/35%, respectively) was strongly shifted toward the U2 sites-derived coding transcript in TALS patients (~15%/85%) in both cell types, hence probably increasing the abundance of the functional full-length protein.

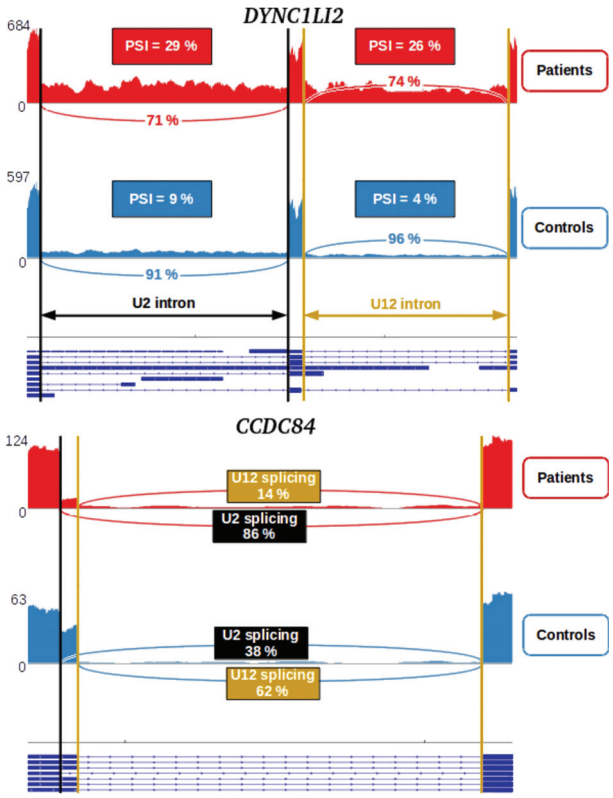

FIGURE 3. Alternative splicing of U12-type introns in TALS patients' fibroblasts. Sashimi plots showing a U2-type intron/U12-type intron coupled retention in the *DYNC1LI2* gene (*top*) and a minor/major spliceosome switching event in the *CCDC84* gene (*bottom*). The *y*-axis corresponds to the mean coverage of each base of the genomic coordinates (*x*-axis). Reference annotations are given on the lowest part of the figure, with annotated exons and introns shown as thick and thin horizontal lines, respectively. U12 and U2 splice sites are marked with yellow and black vertical bars, respectively. Splice junction reads are drawn as arcs connecting a pair of exons. Mean percentage of reads supporting the splicing of either the U12- or U2-type intron are indicated in yellow and black boxes, respectively.

Splicing efficiency of U12-type introns in LCL derived from a TALS patient

The PCA analysis of PSI values obtained with our RNA-seq study showed that in TALS patients, the pattern of U12-type IR in the LCL markedly differed from that seen in fibroblasts and amniocytes (Fig. 1, bottom left). Analysis of the data sets from two TALS LCL technical replicates revealed that U2-type intron splicing was globally unaffected (mean PSI: 4.9% vs. 4.8% for the patient and the control, respectively), while U12-type intron splicing was severely affected. Indeed, the mean ΔPSI obtained from TALS patient and control data sets was 19.7% in the LCL, compared to 4.4% and 3.1% in fibroblasts and amniocytes, respectively (Table 2). Looking into further details, we found that 98% of the 480 U12-type introns with a sufficient number of reads for the analysis were more retained in the TALS than in the control LCL sample and that, strikingly, 79% (370/468) of these retentions had a ΔPSI ≥ 10%, as seen when comparing Figure 4A with Figure 2. Other types of U12-type intron alternative splicing were also far more frequent [69 U12/U2 splice site switching vs. 4 (F) and 2 (A)]. On the other hand, a high level of U2-type IR was not observed, ruling out a sequencing or sample preparation problem. The high magnitude of the U12-type intron splicing defects observed in the LCL of the TALS patient was also unlikely to be due to individual particularities because the comparison in this patient of the mean PSI values obtained for U12-type introns in the LCL versus the fibroblast data sets revealed a marked difference (Supplemental Fig. S6, right). We also found more adjacent U12- and U2-type IRs [18 vs. 9 (F) and 3 (A)], among which two, in *DYNC1LI2* and *DERL2*, were common to all cell type data sets.

The high U12-type IR observed in the present work in the TALS LCL were reminiscent of the massive deregulation of U12-type intron splicing reported in the

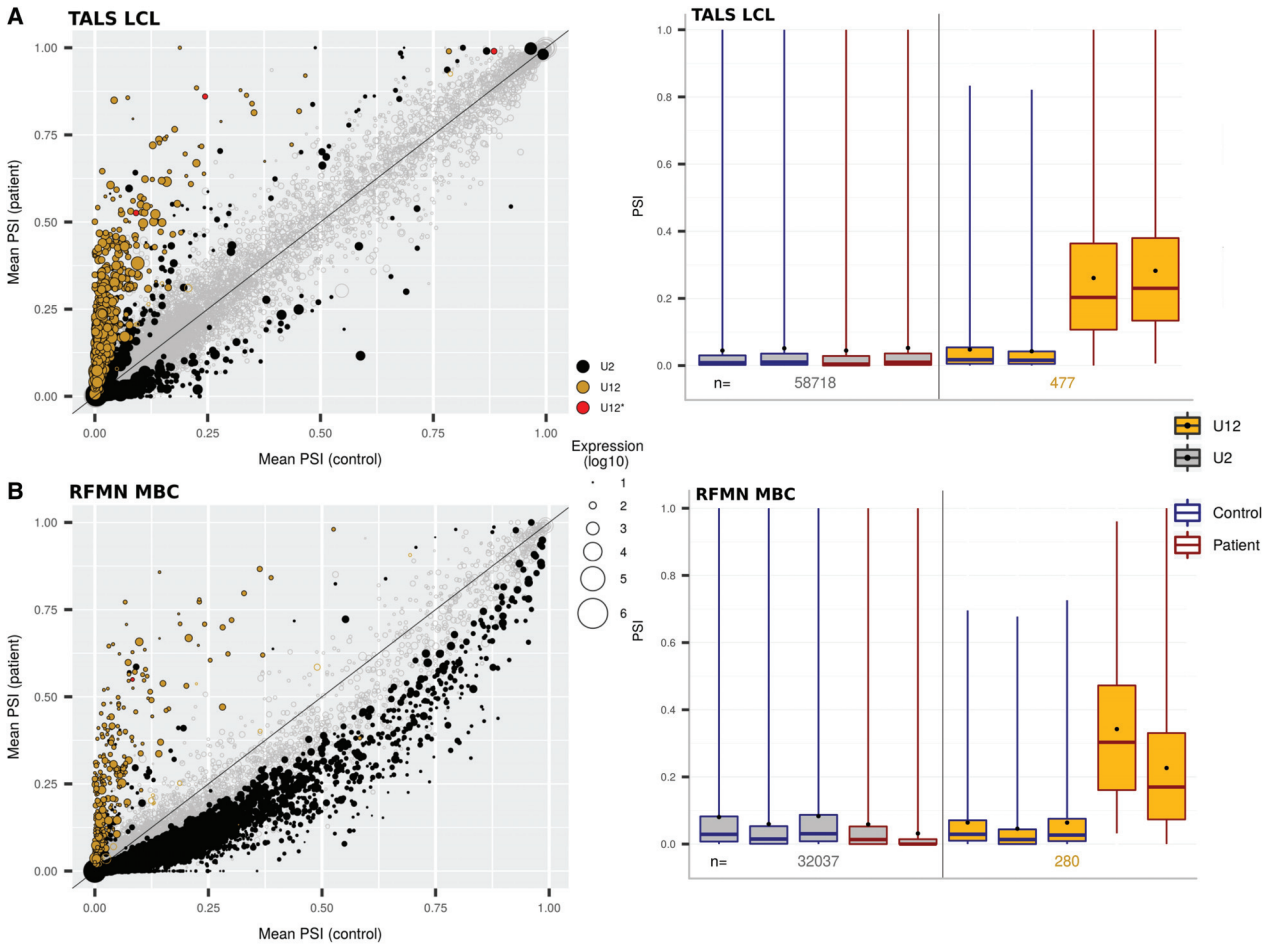

FIGURE 4. Comparison of U12-type IR levels in TALS and RFMN patient and control blood cells. Analysis of the (*A*) TALS LCL (lymphoblastoid cell line) data sets or (*B*) RFMN MBC (mononuclear blood cells) data sets. The TALS patient's and control's LCL data sets consist of two technical replicates for each. (*Left* panels) U12-type intron PSI-plots obtained for the patients' versus the controls' data sets. (*Right* panels) U12-type intron PSI-boxplots of each patient's and control's data set. Legend as in Figure 2.

transcriptomes of blood cells derived from six RFMN patients belonging to six families (Merico et al. 2015; Dinur Schejter et al. 2017; Heremans et al. 2018). In order to investigate further the extent of similarity of U12-type intron splicing patterns in blood cells derived from patients with these two *RNU4ATAC*-associated pathologies, we reanalyzed with the pipelines that we set-up for our study the raw data of the transcriptomic sequences of mononuclear blood cells (MBC) taken from two unrelated RFMN patients (Merico et al. 2015), along with that of three of their heterozygous unaffected relatives (brothers or father of the patients). Different expression and splicing profiles were expected as TALS and RFMN are distinct pathologies and, besides the cells' common tissue's origin (blood), MBC and LCL have marked differences, that is, all types of mononuclear blood cells are present in MBC, while LCL consists of B lymphocytes only, furthermore immortalized by EBV infection, which has been shown to impact gene expression (Lopes-Ramos et al. 2017). Besides, the age at which the blood samples were obtained widely differs between the two studies (babies vs. adults) and the TALS patient and her control are female while the RFMN patients and their controls are male (Table 1). Finally, another important difference was that the TALS data sets were sequenced with higher depth compared to RFMN data sets (125 vs. 47 million of mean aligned reads).

Indeed, not surprisingly given their specificities, PCA showed that U2 and U12 gene expression levels clearly distinguished LCL from MBC; patients and controls from the same collection of data sets grouped together related to the first axis, which explains in both cases more than 65% of the variance (Supplemental Fig. S7A). Concerning U2-type IR, PCA of the mean PSI values did not separate LCL from MBC samples, but separated four of the five MBC samples from the LCL and the fifth MBC samples on the first axis (60% of the variance, Supplemental Fig. S7B, left). These four MBC samples derived from the oldest studied individuals (38, 43, 57, and 67 yr old, compared to 2 mo: LCL sample and 21 yr old: fifth MBC sample), suggesting that age may have an impact on the extent of U2-type IR in blood cells, as previously suggested in the brain (Mazin et al. 2013). Accordingly, we found more than 2000 U2-type IR in the older controls compared to the younger RFMN patients (Fig. 4B, black dots).

Concerning U12-type IR, PCA of the mean PSI values separated TALS and RFMN patients from controls on the first axis (79% of the variance, Supplemental Fig. S7B, right). We did observe separation between TALS LCL and RFMN MBC on the second axis of the PCA, but it explained only 10% of the total variance. When looking at mean U12-type IR values, we observed a strong similarity between the two data sets, as illustrated in Figure 4 (left, yellow dots). Mean PSI were 28.7% in RFMN MBC and 6.0% in control MBC compared to 27.5% and 4.8% in the TALS LCL study, respectively (Table 2), and the mean ΔPSI was 28.9% compared to 27.6%. Because of cell-type specificities and/or different sequencing depths between them, 140 marked U12-type IR found in TALS LCL could not be analyzed in RFMN MBC (13 reciprocally). After filtering them out, we found that 171 marked U12-type IR were common to TALS LCL and RFMN MBC samples (representing 74% and 87% of them, respectively). Of note, only one alternative U12-type intron splicing event, the splice site switching in the uncharacterized *CCDC84* gene, had high and similar ΔPSI in all the patient data sets (TALS fibroblasts, amniocytes, LCL and RFMN MBC, mean $|\Delta PSI| = 54\%$). Altogether, our results suggest that the magnitude of U12-type intron splicing dysfunction could be, firstly, quite similar in blood cells from TALS and RFMN patients, and secondly, highly tissue-dependent, trends that will need to be investigated further.

qRT-PCR validation of U12-type intron missplicing

To confirm the RNA-seq results, we determined the level of retention of nine U12-type introns with various statistically significant mean ΔPSI values ranging from 0 to 37% using a quantitative RT-PCR (qRT-PCR) approach on RNA extracted from fibroblasts derived from five patients and five age- and sex-matched controls. We found a strong concordance between RNA-seq and qRT-PCR mean ΔPSI values using the same metrics ($r^2 = 0.86$, Fig. 5). Of note, even weak effects (mean ΔPSI = 6%) could be confirmed by qRT-PCR.

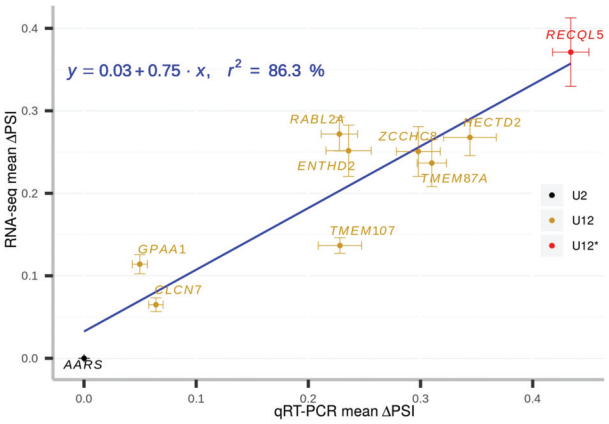

FIGURE 5. Comparison of U12-type IR levels measured by qRT-PCR and RNA-seq. Correlation between the mean ΔPSI values obtained by qRT-PCR when testing introns from ten genes in fibroblasts derived from five patients and their age- and sex-matched controls and those obtained by RNA-seq. Error bars represent standard errors of the mean in both experiments (vertical: RNA-seq; horizontal: qRT-PCR). The linear regression is shown, together with the squared correlation coefficient. The names of the genes whose intron was tested are indicated. The color of gene names indicates the intron type (U12*: U2-type intron proposed to be reclassified as U12-type in this study).

Gene pathways affected in cells derived from TALS patients

Because 97% of the U12-type introns were retained in the TALS LCL data set, precluding classical enrichment analysis, we focused our attention on identifying genes and pathways impacted by U12-type intron missplicing on TALS fibroblasts and amniocytes. As a preliminary study, we first scrutinized the 26 genes with marked U12-type IR common to both data sets (Supplemental Table S1), and found that a high proportion of them were involved in signal transduction (11/26), notably through Notch (*C3orf17*) or Sonic Hedgehog (*IFT22*, *TMEM107*) signaling pathways; genes involved in protein degradation were also represented in a substantial proportion (6/26). We next wanted to look into more details at U12 genes with misspliced transcripts, potentially leading to reduced level of functional proteins, taking into account all the statistically significant differential U12-type IR and U12/U2 splice sites switching found in the two cell types. Toward this goal, we performed a Gene Ontology (GO) term analysis with TopGO (Alexa and Rahnenfuhrer 2016) and compared misspliced to correctly spliced U12 genes, using either the FDR or ΔPSI values as weights. These two analyses revealed 34 and 12 enriched terms, respectively, and we found, in both of them, instances related to developmental processes, response to stimulus, signaling and interestingly, immune system processes (Supplemental Table S5).

DISCUSSION

Transcriptome analysis by RNA-seq has tremendously enhanced our knowledge on gene expression and intron splicing, shedding light on alternative splicing at a large scale and on its relevance in various cellular contexts. However, this technological revolution has mostly benefited to the understanding of U2-type intron splicing. On the other hand, the U12-type introns and U12 genes, as very small minorities, have been largely neglected, despite their acknowledged importance in embryonic development and survival. The few published analyses focusing on U12-type intron splicing were conducted in plants (Gault et al. 2017), fish (Markmiller et al. 2014), or human cancer cells in order to study gene expression regulation (Younis et al. 2013; Niemela et al. 2014). A few additional studies were conducted in the context of pathologies associated with a minor splicing defect, either due to mutations in snRNA components of the minor spliceosome, mainly RFMN syndrome (Merico et al. 2015; Dinur Schejter et al. 2017; Heremans et al. 2018) and early onset autosomal recessive cerebellar ataxia (Elsaid et al. 2017), or in protein components specific to the minor spliceosome, such as observed in isolated familial growth hormone deficiency (Argente et al. 2014), and myelodysplastic syndrome (Madan et al. 2015). Actually, little is known about global U12 gene expression and U12-type intron splicing in physiological conditions in human cells. Therefore, we started our study by tackling these questions in our control data sets consisting of eight fibroblast, four amniocyte, and one lymphoblastoid cell line (LCL) samples derived from control fetuses and children. In these control cells, we found that (i) ∼60% of the 699 U12 genes are consistently expressed in the three different cell types, and (ii) the distribution of the levels of transcriptional expression of U12 genes is highly homogenous between these cell types and peaks at around 30 TPM, as observed for U2 genes. We also observed several occurrences of U12/U2 splice site switching. Alternative splicing of U12-type introns using U2 cryptic donor and acceptor sites, originally described in insects (for review, see Hafez and Hausner 2015), had already been reported in human cells as the result of U6atac snRNA inactivation (Younis et al. 2013), knockdown of the 48K protein (Turunen et al. 2008), and in the context of isolated familial growth hormone deficiency (Argente et al. 2014), and myelodysplastic syndrome (Madan et al. 2015). However, this is the first time that such alternative splicing events are found to occur physiologically in humans. Because the consensus sequences for the acceptor sites of U2- and U12-type introns are less divergent than that of the donor sites, we suppose that the major spliceosome was used for splicing the U2 donor-U12 acceptor mixed introns and the minor spliceosome for the less abundant U12 donor-U2 acceptor mixed ones.

After having determined the frame of U12 gene expression and U12-type intron splicing in the context of a functional minor spliceosome, we set out to identify the consequences of biallelic *RNU4ATAC* mutations within these cell types in five fibroblast, three amniocyte and one LCL samples derived from seven unrelated TALS patients. Rather surprisingly, we did not observe any impact of such mutations on U12 or U2 gene expression in fibroblasts or amniocytes derived from TALS patients, although we used the tool (DESeq2) and cutoffs [FDR $\leq$ 5%; $|\log_2(FC)| \geq 1$] recommended for such a data set size (five patients vs. eight controls) (Schurch et al. 2016). Our previous qRT-PCR study on fibroblasts derived from two homozygous g.51G > A TALS patients and two age- and sex-matched controls (biosamples also included in the present study) had shown that 12 of the 22 tested U12 genes—chosen randomly among those reported as being expressed in the skin—presented a differential expression (Edery et al. 2011). However, we now believe that this previous result most likely stemmed from biological and/or inter-individual variations that could not be correctly modeled due to the small number of samples, and illustrate the necessity to use more stringent criteria when studying a very small number of biological samples. Although we cannot rule out the possibility that a number of U12 genes may

be slightly differentially expressed—but identifying them would require more than 20 biological replicates (Schurch et al. 2016)—we conclude that U12 gene expression levels, that is, the number of poly(A)$^+$ transcripts produced, are essentially unchanged in TALS fibroblasts and amniocytes compared to controls.

Then, we studied splicing efficiencies and found that most U12-type introns were significantly retained in the TALS transcriptomes whatever the cell type studied. Hence, even though the number of poly(A)$^+$ transcripts is unchanged for most genes in patients' compared to controls' cells, a fraction of them, larger in patients, contains U12-type IR and cannot lead to functional proteins. Although these IR were statistically significant, we found that their magnitudes were small in the fibroblast and amniocyte data sets, with only 14% and 18% of the retained U12-type introns showing a $\Delta PSI \geq 10\%$, respectively. In contrast, these U12-type IR were much more pronounced in the LCL data set, as 79% had a $\Delta PSI \geq 10\%$. Considering that the overall transcript levels are unchanged but splicing is altered, we conclude that the number of transcripts that could be translated into functional proteins is therefore mildly decreased in fibroblasts and amniocytes, and largely decreased in lymphocytes. The extreme rarity of the TALS syndrome and the premature death of the children affected with this disease did not permit us to collect additional blood samples up to now and hence analyze LCL biological replicates. Nevertheless, several lines of evidence support the assumption that peripheral blood cells may exhibit particularly pronounced U12-type IR: (i) This difference in U12-type intron splicing efficiency was clearly visible when comparing cells derived from skin and blood taken the same day on the same TALS child, while no difference was seen for another TALS child between amniotic fluid taken in utero and skin at 10 mo of age (Supplemental Fig. S6); (ii) similar high levels of U12-type IR were observed in the RFMN MBC and TALS LCL data sets, despite the different pathologies, blood cell subtypes analyzed, gender and age of the patients, and RNA-seq settings (Fig. 4); (iii) a comprehensive analysis of IR performed on 52 human samples from different cell and tissue types showed that the highest percentage of retention was found in white blood cells (>30%, compared to <5% in fibroblasts) (Braunschweig et al. 2014).

We observed that the competition between the major and minor spliceosomes for splicing some introns, which we show here for the first time to occur physiologically in humans, is more favorable to the major spliceosome in TALS amniocytes, fibroblasts and LCL as compared to the situation seen in control cells. This was particularly pronounced for the *CCDC84* gene, thereby increasing the amount of the full length protein of as yet uncharacterized function.

Unexpectedly, exclusively in the TALS LCL data set (Supplemental Fig. S9), we found reads for all spliceosomal snRNAs at the exception of U6 and U6atac, an observation also made in a previous analysis of RFMN data sets (Dinur Schejter et al. 2017). This was unexpected because snRNAs belong to the nonpolyadenylated class of RNAs, yet we performed RNA-seq experiments on poly(A)$^+$ RNAs. We postulate that the accumulation of polyadenylated snRNA precursors may have resulted from a deficient Integrator complex, which plays a pivotal role in the 3′-end processing of the snRNAs transcribed by RNA Polymerase II, that is, all snRNAs apart from U6 and U6atac (for review, see Guiro and Murphy 2017). Integrator contains at least 14 subunits, of which four are encoded by U12 genes, namely *INTS4*, *INTS7*, *INTS8*, and *INTS10*, markedly differentially mis-spliced in TALS LCL ($\Delta PSI_{INTS4} = 11.4\%$; $\Delta PSI_{INTS7} = 28.6\%$; $\Delta PSI_{INTS8} = 16.5\%$; $\Delta PSI_{INTS10} = 34.1\%$). In contrast, the U12-type introns of the three U12 Integrator genes expressed in the TALS fibroblast data sets had a very low ΔPSI value ($\Delta PSI_{INTS7} = 2.5\%$; $\Delta PSI_{INTS8} = 1.6\%$; $\Delta PSI_{INTS10} = 6.3\%$). Interestingly, mutations in *INTS1* and *INTS8* are associated with impaired RNA splicing in rare recessive neurodevelopmental syndromes with developmental delay and distinctive appearance (Oegema et al. 2017). However, the absence of massive U2-type intron splicing defects in LCL attests that despite this maturation default, the amount of functional snRNAs of the major spliceosome is sufficient for efficient U2-type intron splicing and that U12 Integrator genes missplicing is unlikely to be the primary cause of the high magnitude of U12-type intron missplicing in this LCL sample.

We observed that the level of IR is quite variable among U12-type introns, even in TALS fibroblasts and amniocytes where most introns are retained in only a marginal fraction of transcripts. To try to understand why some U12-type introns are more sensitive to a defective spliceosome than others, we considered a number of intron features previously shown to influence IR in mammals, for example, donor/acceptor splice site scores, GC content, intron length (Braunschweig et al. 2014), and correlated them with the level of U12-type intron missplicing using a linear model (Supplemental Table S6). Among the many features tested, only two were found to significantly correlate with PSI values in patients. The first one is the PSI value in controls (50% of the variance), which means that introns poorly spliced in controls are even more poorly spliced in patients. The second one is the gene expression level (10% of the variance), which means that poorly expressed genes are more subject to missplicing than the more expressed ones, as had been previously reported for U2-type introns (Saudemont et al. 2017). We also searched for enriched motifs such as splicing enhancers that might bind a splicing factor (Dietrich et al. 2001b) for explaining high PSI values but we were unable to identify such sequences, leaving open the question of the remaining features causing U12-type intron "ultra sensitivity" to a defective spliceosome for some of them.

Transcriptome analyses show much promise in elucidating the pathogenesis of genetic diseases, even more in those due to a splicing defect. However, it is well known that expression programs for genes involved in development are highly time- and tissue-specific, and even cell-specific in the early stages of embryogenesis. The understanding of the molecular mechanisms involved in the pathogenesis of TALS will require additional transcriptomic analyses to be performed on different cell types at various developmental or differentiation stages, hence necessitating to generate induced pluripotent stem cells and/or develop animal models. Nevertheless, the present finding that TALS and RFMN blood cells share a similar pattern of U12-type IR and that the GO term analysis performed on the TALS fibroblast and amniocyte data sets showed an enrichment in immunity-linked terms suggest that thorough investigation of TALS immune phenotype should be carried out.

MATERIALS AND METHODS

Identification of U12-type introns in the human genome

U12DB, the U12 Intron Database (http://genome.crg.es/cgi-bin/u12db/u12db.cgi) released in 2006 by T. Alioto with the aim to catalog U12-type introns of completely sequenced eukaryotic genomes (Alioto 2007), has not been updated since its launching. We updated the list of human U12-type introns using T. Alioto's scoring matrices on a more recent genome annotation [Gencode v19 (GRCh37)/Ensembl v75], the latest one at the time of the analysis of the pilot project data sets. Out of 289,023 introns annotated in Gencode v19, the pipeline classifies 846 of them as U12-type introns (Supplemental Table S3). Those are located in 699 genes, of whom 105, 20, 3, and 1, respectively contained 2, 3, 4, and 5 U12-type introns. When more than one U12-type intron is present in a gene, in most cases (85/129), the coordinates of at least two of these U12-type introns overlap, indicating that the same U12-type intron can be spliced out using alternative U12 consensus splice sites.

Biological samples

Biospecimens were obtained from seven unrelated TALS cases, four children (three *RNU4ATAC* homozygotes and one *RNU4ATAC* compound heterozygote) and three fetuses (one *RNU4ATAC* homozygote and two *RNU4ATAC* compound heterozygotes), and deposited to the Lyon University Hospital Biobank dedicated to genetic diseases for processing, storage and management (CBC Biotec of the Hospices Civils de Lyon, certified with a specific French standard for biobanks, NF S96-900). These biospecimens consisted of skin biopsies and amniotic fluid from which primary fibroblasts and amniocytes were respectively derived, and peripheral blood from which lymphoblastoid cell lines (LCL) were established, following standard procedures. For two children, two different types of samples were obtained: peripheral blood and skin biopsy for one, amniotic fluid during gestation and skin biopsy after birth for the other. Adequate biological samples from age- and sex-matched controls were provided by the CBC Biotec biobank. Informed written consent for the use of these samples in research was obtained from all parents of TALS patients, TALS fetuses and control fetuses and children. The detailed characteristics of the analyzed samples, including the information on whether they derived from post-mortem material, are described in Table 1.

RNA extraction

RNA extractions were performed using the Nucleospin RNA kit (Macherey Nagel) according to the manufacturer's recommendations. A further round of DNase (Promega) treatment was systematically performed to remove any possible residual amount of DNA. Total RNA concentration was then quantified with a NanoDrop spectrophotometer (Nanodrop Technologies) and RNA quality assessed using the Agilent 2100 Bioanalyzer (Agilent). RNA integrity number (RIN) was >7 in all cases.

cDNA library preparation, high-throughput sequencing

One to two micrograms of RNA were sent for RNA-sequencing to IntegraGen Genomics (Evry, France), where a DNA library was generated with the "TruSeq Stranded mRNA Sample Prep" kit (Illumina) that comprises a step of mRNA purification using oligo(dT) beads. A total of 28 RNA-seq experiments have been performed at two different times: (i) A pilot study was performed on a HiSeq 2000 sequencer (Illumina), yielding approximately 716 million of nonstranded two time 100 bp paired-end reads, with librairies obtained with RNA extracted from skin fibroblasts taken on two TALS children homozygous for g.51G > A and from the LCL derived from one of these children, and from their matched controls (six RNA-seq experiments, see Table 1). The reads thus obtained were analyzed as described in the following paragraph: it showed that the extent of IR being low, additional samples needed to be analyzed in order to obtain reliable results. (ii) An extended study was later performed on a HiSeq 4000 sequencer (Illumina), yielding approximately 2670 million of stranded two time 75 bp paired-end reads, with librairies obtained with RNA extracted from all the samples, including those already sequenced in the pilot study in order to have technical replicates for some of them (22 RNA-seq experiments, see Table 1). Sequencing metrics are given for each sample in Supplemental Table S7. Raw RNA-seq data are available upon request.

qRT-PCR

cDNA synthesis was carried out with 1 µg DNA-free RNA (the same batches as those used for RNA-seq) using GoScript Reverse Transcription System and oligodT primers (Promega) according to the manufacturer's protocol. qRT-PCR was performed using the Rotor-Gene SYBR Green PCR kit and Rotor-Gene Q (Qiagen) according to the manufacturer's protocol. All experiments were done in three replicates.

Bioinformatics analysis of RNA-seq data

Splicing analyses

Our three bioinformatics pipelines, shown in Supplemental Figure S10, are composed of multiple steps executed by various tools to achieve three goals: (i) read alignment/assembly; (ii) read quantification; (iii) alternative splicing event quantification and statistical analysis, with a special focus on IR.

IR identification and quantification in RNA-seq data is a difficult bioinformatics task for multiple reasons (discussed in detail in Vanichkina et al. 2018). To date, only four dedicated tools are available: vast-tools (Tapial et al. 2017) (which can also detect other types of alternative splicing events); IRcall/IRclassifier (Bai et al. 2015); intEREst (Oghabian et al. 2018), and IRFinder (Middleton et al. 2017). Their main difference lies in the intronic read quantification method: Vast-tools outputs the number of exon-intron junctions reads, IRcall/IRclassifier the number of reads aligned to the full intron, intEREst and IRFinder the read coverage of specific intronic regions that do not correspond to low complexity regions or alternative exons, hence improving precision. Furthermore, IRFinder reduces the impact of heterogeneous coverage by discarding 60% of the intronic regions' bases containing the highest and lowest covered bases, and it also outputs the number of exon–intron junctions reads. We thus chose to use IRFinder, being the most precise tool.

IR detection and quantification method (IRFinder v1.2.0, mapping-first)

RNA-seq read quality control was performed using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped with STAR v2.5.0b (Dobin et al. 2013) using IRFinder's custom STAR index of the latest annotation of the GRCh37 assembly (Ensembl v75) and default parameters.

Splicing-supporting reads are reads aligning to an exon-exon junction. An overhang (minimal number of bases around a junction covered by an aligned read) of five was required to consider that any read is aligned to a junction. Hence, the number of unique positions on a junction where a 75 bp read (in the case of our study) can be aligned (referred to as "effective size") is:

$$\text{effective size} = \text{read length} - 2 \times \text{overhang} + 1$$

$$\text{effective size} = 66.$$

Any read fully covering one of these 66 positions will be counted as a splicing-supporting read.

We then used two strategies to select splicing- and retention-supporting reads depending on the number of an intron's informative bases.

Retention-supporting reads can either be reads aligning to the intron body or reads aligning to one of the two exon-intron junctions. In general, IRFinder will use both intron body reads and exon-intron reads. In the cases where the number of informative bases is too low (≤40 bases or <70% of the total number of bases of the intron, 31% of U2- and U12-type introns), for example, most of the intron length is covered by repeats or annotated features, IRFinder is conservative and reasonably chooses not to compute the intronic read coverage. However, it also does not compute exon–intron junction reads coverage. We argue that this latter quantification could still be of interest, as although it does not indicate the full intronic coverage, it still testifies that this peculiar intron's splice sites were not used for splicing. In other words, exon–intron junctions quantification does not indicate the type of the alternative splicing event (it could either be an IR or a alternative donor + acceptor), but still indicates the amount of unspliced intron. In order to force IRFinder to do the exon–intron quantification for all introns, we rewrote a specific test in its code (*intronExclusion.pl*, line 83: *if ($newlen > 40 && ($newlen/$len) ≥ 0.7) {* replaced by *if ($newlen > 0) }*). For the special case of U12-type introns with no informative base (167 cases), IRFinder 1.2.0 could not be run and we had to develop a custom python script (*junctionsCover2IRF.py*) to do the quantification. For a given list of introns, a read length and BAM files, this script uses samtools view to quantify the number of aligned reads on the exon–exon junction and the two exon–intron junctions and creates a file formatted in the same way as a conventional IRFinder file, allowing us to merge them together (Supplemental Fig. S10). In the following analyses, the retention-supporting reads will either be reads aligned to the intron body, or reads aligned to the exon–intron junctions if the number of informative bases of the intron was smaller than the effective length of the exon–intron junction (66 nt).

The list of introns we analyze corresponds to constitutively spliced introns, but also to alternatively spliced introns, some of which are spliced out only in specific tissues. Out of all introns analyzed, some are never seen spliced out in our data sets. We chose not to consider them as introns as they would otherwise artificially increase IR rates. In practice, we did not analyze IR for all introns with less than five splicing-supported reads on average in the control samples and for each cell-type. Among the IR sufficiently covered (see Filters for PSI and ΔPSI computation below), this filters out 4687, 5468, 4469, and 3239 introns (among which 12, 13, 18, and 7 were U12-type introns) in the primary fibroblast, amniocyte, LCL, and MBC data sets, respectively.

Alternative splicing events detection with annotation (vast-tools v2.0.0, mapping-first)

In addition to IR, vast-tools (Tapial et al. 2017) can also detect three other types of alternative splicing events: alternative donor, alternative acceptor and exon skipping. Vast-tools results concerning IR are not presented because >99% of the differential U12-type IR detected by vast-tools were also found by IRFinder, but only 35% or less of the differential U12-type IR detected by IRFinder were also found by vast-tools. All results are however available in the supporting shiny interface (http://lbbe-shiny.univ-lyon1.fr/TALS-RNAseq/). Briefly, vast-tools aligns the reads from each sample on different references (genome, exon–exon junction, …) using BOWTIE (v1.1.2.), and then analyzes the alignment file to quantify the number of reads supporting the inclusion or exclusion of an exon for each of its 213,087 possible alternative splicing events annotated in its database.

Some introns from the vast-tools' splicing event database were not annotated in Ensembl v75. In order to determine their type, we ran T. Alioto's scripts on these introns if both of their splice sites were annotated in Ensembl v75. This resulted in 56 new U12-type introns (of which 55 overlapped with a known U12-type intron but used a different acceptor site, Supplemental Table S3).

Because this method cannot detect alternative splicing events which are absent from its database, we also used an assembly-first

and annotation-free method for alternative splicing events detection.

Alternative splicing events detection without annotation (KisSplice v2.4.0-p1, assembly-first)

Briefly, KisSplice (Sacomoto et al. 2012) assembles the reads in a de Bruijn graph and searches for so-called bubbles in this graph, which correspond to alternative splicing events. The two paths of the bubble are then mapped to a reference genome using STARlong v2.5.0b, and the resulting alignments are processed by KisSplice2RefGenome to annotate the event, by assigning it notably a gene name and an AS subtype. We recently showed (Benoit-Pilven et al. 2018a) that this assembly-first approach was particularly adapted to identify novel splice sites. This advantage comes at the expense of poorer performance for long and unfrequent variants, because de novo assembly requires more coverage. This is the reason why we do not use KisSplice to analyze IR.

Counts normalization, PSI/ΔPSI computation, and differential analysis (kissDE)

IRFinder, vast-tools and KisSplice all output the number of reads supporting the inclusion (i.e., a retained intron or exon) or exclusion (i.e., a spliced intron or skipped exon) transcript in each sample and for each IR/alternative splicing event. The bioconductor R package kissDE (https://www.bioconductor.org/packages/devel/bioc/html/kissDE.html) (Benoit-Pilven et al. 2018b) was then used for counts normalization, splicing event strength estimation (PSI and ΔPSI) and differential analysis between two conditions (FDR).

Briefly, kissDE starts by normalizing the read counts for the library sizes using DESeq2, and by normalizing the inclusion-supporting reads by the length of the inclusion (that can be very large for IR events) compared to the exclusion. Then, for each splicing event, kissDE computes a PSI for each sample. In the context of this study, where several replicates are available for the patients and controls, a mean PSI is calculated for each condition, and corresponds to the patients/controls PSI used throughout this article. In the Results section, we also used the mean ΔPSI of all U12- or U2-type introns in a data set. Finally, a differential analysis is run that models counts with either a Poisson (for technical replicates) or a negative binomial (for biological replicates) distribution, and uses the generalized linear model framework to model the expected signal intensity. A likelihood ratio test is used to estimate the probability of an interaction between the splice-forms (inclusion and exclusion) and the condition. The Benjamini–Hochberg procedure is used to account for multiple testing and compute FDR values.

We considered an alternative splicing event statistically significant if its FDR ≤ 5%, and markedly significant if, in addition, its |ΔPSI| ≥ 10%.

Local expression value

The local expression (*locExp*) value, calculated for each intron, is the number of reads attesting to either the inclusion or exclusion of an intron, and is defined as:

$$locExp = excReads + incReads/2$$

$$locExp^* = excReads^* + incReads^*/2,$$

with *excReads* the number of reads on the exon–exon junction and *incReads* the number of reads on both exon–intron junctions. A star indicates library-sized normalized counts.

The main advantage of using the local expression value is that there is no need to infer full-length transcripts and their abundance, a notoriously difficult and error-prone task (Steijger et al. 2013), to derive an estimation of transcripts expression. It also has the advantage of directly focusing on transcripts which contain the exons flanking the intron of interest. In contrast, a measure of gene expression based on counting all reads falling within the gene boundaries will also include reads stemming from transcripts which do not overlap the intron of interest, for instance in the case of alternative transcription start/end. It will also be confronted with the difficult task of correctly estimating gene length, in the presence of multiple alternative transcripts.

Filters for PSI and ΔPSI computation

To compute robust metrics, we apply a coverage threshold on the local expression of an alternative splicing event. In a sample, both the local expression and the normalized local expression values must be ≥10 to compute the PSI value of an intron. At least half of the patients and half of the controls must have a computed PSI in order to have a ΔPSI estimation.

Differential gene expression analysis method (DESeq2)

We tested if genes were differentially expressed between our two conditions with the DESeq2 conventional pipeline (Love et al. 2014) HTSeq tool to generate gene expression values (Anders et al. 2015).

Principal component analyses (PCA)

We used the *dudi.pca* function from the R package ade4 v1.7-11 (https://github.com/sdray/ade4) (Bougeard and Dray 2018) on either a table of TPM or PSI. For each PCA, the most variable values (up to 500) were used (as conventionally done in DESeq2) and the first (PC1) and second (PC2) most explanatory axes were plotted. We compared the percentage of the variance explained by each axis of these PCA (*PCAvar*) to the mean of the ones obtained after randomizing independently each row of the TPM or PSI table 100 times (*randomVar*). Axes with explained variance smaller or equal to our randomized data (*PCAvar* ≤ *randomVar*) are denoted with *ns* (not significant) and should not be interpreted.

Intron retention validations

IR validations were carried out with RNA extracted from fibroblast cell lines derived from patients TALS2, TALS4, TALS6 (all g.51G > A homozygous), TALS10 (g.50G > C;g.51G > A) and TALS3 (g.40C > T;g.124G > A) and from five control children or fetuses matched for age and gender. We tested introns with various extent of IR (i.e., mean ΔPSI) from eight U12 genes (*CLCN7*: 6.5%, *GPAA1*: 11.4%, *TMEM107*: 13.7%, *TMEM87A*: 23.7%, *ZCCHC8*: 25.1%, *ENTHD2*: 25.2%, *HECTD2*: 26.8%, *RABL2A*: 27.2%), one U2 gene reclassified as U12 in this study (U12*, *RECQL5*: 37.1%) and one control U2 gene that did not display

IR (*AARS*: 0%). The *ACTB* gene (encoding β actin) was chosen as the endogenous control. To be able to compare IR measured by qRT-PCR to that measured by RNA-seq, we computed the mean ΔPSI for each gene from qRT-PCR experiments as follows:

$$Rq_{i,t,C} = 2^{C^*t_C - Ct_{i,t,C}},$$

$$PSI_{i,C} = \frac{Rq_{i,r,C}}{Rq_{i,r,C} + Rq_{i,s,C}},$$

$$PSI_{Ctrl} = \frac{\sum_{i=1}^{i=3} PSI_{i,Ctrl}}{3},$$

$$\Delta PSI_i = PSI_{i,Patient} - PSI_{Ctrl},$$

$$\Delta PSI = \frac{\sum_{i=1}^{i=3} \Delta PSI_i}{3},$$

with *Ct* the number of qRT-PCR cycle needed for the fluorescence to cross a given threshold (125), * denoting the mean Ct (from the three technical replicates) of the endogenous control (*ACTB*), *i* the technical replicate, *t* the type of transcript quantified (either *r* or *s* for transcript retaining or splicing the intron), *C* the experimental conditions (either *Ctrl* or *Patient*) and *Rq* the relative quantification of the DNA with respect to the endogenous control. The RNA-seq mean ΔPSI was computed for each gene by subtracting the PSI of each matched control/patient pair, and calculating the mean.

GO terms enrichment analysis

We searched for GO terms enriched in our set of genes with U12-type differentially spliced introns to highlight potential biological processes specifically disrupted in patients and thus, possibly related to the phenotype. We used the topGO (Alexa and Rahnenfuhrer 2016) (v2.30.1) R package using the genes for which a U12-type intron alternative splicing event had been tested and a user provided quantitative score for each gene. We performed two different analyses using distinct scores. The first one defined the score based on the FDR of a gene's U12-type intron (minimum FDR in the special case where a gene harbors multiple U12-type introns or multiple splicing events for the same U12-type intron), which correspond to the classical use of TopGO. Genes gain more weight as their FDR value is close to 0. This should detect GO terms enriched in genes with the most reproducible U12-type intron alternative splicing events compared to unaffected genes. The second analysis defined the score based on the |ΔPSI| of a gene's U12-type intron: The score is either 0 for genes without any significant U12-type intron alternative splicing event or the |ΔPSI| (maximum |ΔPSI| in the special case described above). Genes gain more weight as their |ΔPSI| is close to 1. This should detect GO terms enriched in genes with the highest differences of missplicing between patients and controls compared to unaffected genes. For each analysis, we used the Kolmogorov–Smirnov test to account for the weights and we reported a GO term as enriched if its *P*-value was ≤5% for either of our two analyses.

In each analysis, the default "weight01" algorithm was used. The following describes the parameters used to create the topGOdata object in R: We searched for biological process (*ontology* = "*BP*") in the Gene Ontology DataBase version from October 2018 (*mapping* = "*org.Hs.eg.db*", *annot* = *annFUN.org*) using Ensembl ID (*ID* = "*ensembl*"). The enriched GO terms were mapped to a subset of more generic GO terms (GO slim) using the GSEABase R package v1.44.0 and the GO slim AGR subset (*go_slim.agr*) downloaded on the GeneOntology website (http://geneontology.org/docs/go-subset-guide/).

Features influencing U12-type IR

In order to identify features that could have an impact on the level of U12-type IR, we used a linear model. We worked on all analyzed introns (using filters described in Materials and Methods) in the fibroblasts data set. We wanted to explain the U12-type introns' mean PSI (mPSI) of the patients with a set of 32 explicative variables (hereafter referred to as predictors), see Supplemental Table S4. We used a log transformation of mPSI (Supplemental Fig. S11A), since diagnostic plots (Supplemental Fig. S11B–E) show that the assumptions of linear regression are much better satisfied with this transformation. Zero values were replaced by the minimum nonzero mPSI divided by two to guarantee that the transformed value for zero is still lower than all other values. For 15/32 predictors, we needed to define a major transcript for each intron. In the case of multiple transcripts, we chose the CCDS, as annotated in APPRIS (Rodriguez et al. 2013). In the case of multiple CCDS, we chose the longest ORF. In case of ties, we chose the longest transcript. We first performed a simple linear regression to test each predictor in an independent way [model = log(mPSI) ~ predictor] using R version 3.5.1 and anova [lm(model)], for the fitting and the variance analysis. *P*-values and *R*-squared (R^2) values (indicating the percentage of variance explained) for each predictor are both reported in the Supplemental Table S4. Then, we ordered the significant (*P*-value ≤ 5%) predictors by decreasing R^2 value (predictor1, predictor2, …, predictorN). From the initial model m0 = log(mPSI) ~ predictor1, we created the multiple linear regression model m∞ = m0 + predictor2. We then compared these two nested models, using a likelihood ratio test [anova(lm(m0), lm(m∞))], to decide whether the additional predictor could be considered as significantly associated to IR. If the *P*-value was ≤5% and $R^2 \geq 1\%$, we set m0 = m∞, else we kept the same unchanged m0. We did this up to predictorN to build the complete model. Each R^2 value is computed by dividing the sum-of-square of each predictor by the sum of the sum-of-square of all predictors. The same analysis was run to explain the U12-type introns' mean PSI of the controls.

Motif sequences analysis

In order to identify motifs enriched in differentially retained U12-type introns compared to other analyzed U12-type introns, we used the MEME Suite 5.0.1. software (Machanick and Bailey 2011; Bailey et al. 2015). Tested U12-type introns were separated into two groups: candidates ($n = 49$), for which a strong differential IR was detected in Patients (FDR ≤ 5% and ΔPSI ≥ 10%), and unchanged ($n = 45$), for which no differential IR was detected in Patients (FDR ≥ 20% and |ΔPSI| < 1%). In the case of overlapping U12-type introns, the largest one was conserved. Sequences of each intron, with 100 bp upstream and downstream the intron, were retrieved with bedtools' fastaFromBed (v2.25.0).

Sequences of introns on the minus strand were reverse complemented. In order to have groups with comparable intron length, we calculated the minimum length ratio of each sequence from the candidates group with each sequence from the unchanged group (minRatioLength) and we selected the sequences with a minRatioLength ≥ 0.95. This step resulted in 34 selected sequences in the candidates group and 39 sequences selected in the unchanged group. With MEME, we searched for ungapped motifs of length 8 to 50. The OOPS (One Occurrence Per Sequence), ZOOPS (Zero or One Occurrence Per Sequence) and ANR (Any Number of Repetitions) mode of MEME were used (-mod oops|zoops|anr) with the "differential enrichment" objective function (-objfun de) to detect motifs significantly enriched either in the candidates or in the unchanged sequences (E-value $\leq 5\%$, -evt 0.05). All other parameters were set to default values. With DREME, we searched for small (up to 8 nt) ungapped motifs differentially enriched in either the candidates or unchanged sequences. The -norc option was used; other parameters were set to default value.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by CNRS, Inserm, Université Paris 7 and Université Lyon 1 through recurrent funding, the ANR Aster (no. ANR-16-CE23-0001) and U4ATAC-BRAIN (no. ANR-18-CE12-0007-01) grants and an Inserm/Hospices Civils de Lyon grant to P.E. (Contrat d'Interface pour Hospitaliers). A.C. was supported by a grant from Inria (Thése Inria-Inserm "Médecine Numérique" - 2016) and C.B.P. by a grant from the Fondation pour la Recherche Médicale to P.E. (Financement d'un ingénieur - ING20160435660). We thank the families for their participation in this study, the CBC Biotec biobank for sample management (Emilie Chopin, Isabelle Rouvet), the students for their contribution (Clément Saccaro, Gabriel Sala, Nabil Sersoub), Integragen SA for performing the RNA-seq experiments, Daniele Merico for critical reading of the manuscript and helpful suggestions, Tyler Alioto for providing us with the scripts for predicting U12 introns and Cyril Bourgeois, Vincent Navratil and Marion Delous for stimulating discussions.

Received March 29, 2019; accepted May 28, 2019.

REFERENCES

Alexa A, Rahnenfuhrer J. 2016. *topGO: enrichment analysis for gene ontology*. R package version 2320.

Alioto TS. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* **35:** D110–D115. doi:10.1093/nar/gkl796

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31:** 166–169. doi:10.1093/bioinformatics/btu638

Argente J, Flores R, Gutiérrez-Arumí A, Verma B, Martos-Moreno GÁ, Cuscó I, Oghabian A, Chowen JA, Frilander MJ, Pérez-Jurado LA. 2014. Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. *EMBO Mol Med* **6:** 299–306. doi:10.1002/emmm.201303573

Bai Y, Ji S, Wang Y. 2015. IRcall and IRclassifier: two methods for flexible detection of intron retention events from RNA-seq data. *BMC Genomics* **16(**Suppl 2**):** S9. doi:10.1186/1471-2164-16-S2-S9

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME suite. *Nucleic Acids Res* **43:** W39–W49. doi:10.1093/nar/gkv416

Benoit-Pilven C, Marchet C, Chautard E, Lima L, Lambert M-P, Sacomoto G, Rey A, Cologne A, Terrone S, Dulaurier L, et al. 2018a. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Sci Rep* **8:** 4307. doi:10.1038/s41598-018-21770-7

Benoit-Pilven C, Marchet C, Kielbassa J, Brinza L, Cologne A, Siberchicot A, Lacroix V. 2018b. *kissDE: retrieves condition-specific variants in RNA-seq data*. R package version 113.

Bogaert DJ, Dullaers M, Kuehn HS, Leroy BP, Niemela JE, De Wilde H, De Schryver S, De Bruyne M, Coppieters F, Lambrecht BN, et al. 2017. Early-onset primary antibody deficiency resembling common variable immunodeficiency challenges the diagnosis of Wiedeman-Steiner and Roifman syndromes. *Sci Rep* **7:** 3702. doi:10.1038/s41598-017-02434-4

Bougeard S, Dray S. 2018. Supervised multiblock analysis in R with the ade4 package. *J Stat Softw* **86:** 1–17. doi:10.18637/jss.v086.i01

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24:** 1774–1786. doi:10.1101/gr.177790.114

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. *Mol Cell* **2:** 773–785. doi:10.1016/S1097-2765(00)80292-0

Chang W-C, Chen Y-C, Lee K-M, Tarn W-Y. 2007. Alternative splicing and bioinformatic analysis of human U12-type introns. *Nucleic Acids Res* **35:** 1833–1841. doi:10.1093/nar/gkm026

Dietrich RC, Peris MJ, Seyboldt AS, Padgett RA. 2001a. Role of the 3′ splice site in U12-dependent intron splicing. *Mol Cell Biol* **21:** 1942–1952. doi:10.1128/MCB.21.6.1942-1952.2001

Dietrich RC, Shukla GC, Fuller JD, Padgett RA. 2001b. Alternative splicing of U12-dependent introns in vivo responds to purine-rich enhancers. *RNA* **7:** 1378–1388.

Dietrich RC, Fuller JD, Padgett RA. 2005. A mutational analysis of U12-dependent splice site dinucleotides. *RNA* **11:** 1430–1440. doi:10.1261/rna.7206305

Dinur Schejter Y, Schejter YD, Ovadia A, Alexandrova R, Thiruvahindrapuram B, Pereira SL, Manson DE, Vincent A, Merico D, Roifman CM. 2017. A homozygous mutation in the stem II domain of *RNU4ATAC* causes typical Roifman syndrome. *NPJ Genom Med* **2:** 23. doi:10.1038/s41525-017-0024-5

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Edery P, Marcaillou C, Sahbatou M, Labalme A, Chastang J, Touraine R, Tubacher E, Senni F, Bober MB, Nampoothiri S, et al. 2011. Association of TALS developmental disorder with defect in minor splicing component *U4atac* snRNA. *Science* **332:** 240–243. doi:10.1126/science.1202205

Elsaid MF, Chalhoub N, Ben-Omran T, Kumar P, Kamel H, Ibrahim K, Mohamoud Y, Al-Dous E, Al-Azwani I, Malek JA, et al. 2017. Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia. *Ann Neurol* **81:** 68–78. doi:10.1002/ana.24826

Farach LS, Little ME, Duker AL, Logan CV, Jackson A, Hecht JT, Bober M. 2018. The expanding phenotype of *RNU4ATAC*

pathogenic variants to Lowry Wood syndrome. *Am J Med Genet A* **176:** 465–469. doi:10.1002/ajmg.a.38581

Ferrell S, Johnson A, Pearson W. 2016. Microcephalic osteodysplastic primordial dwarfism type 1. *BMJ Case Rep* **2016:** Jun 16. doi:10.1136/bcr-2016-215502

Gault CM, Martin F, Mei W, Bai F, Black JB, Barbazuk WB, Settles AM. 2017. Aberrant splicing in maize *rough endosperm3* reveals a conserved role for U12 splicing in eukaryotic multicellular development. *Proc Natl Acad Sci* **114:** E2195–E2204. doi:10.1073/pnas.1616173114

Guiro J, Murphy S. 2017. Regulation of expression of human RNA polymerase II-transcribed snRNA genes. *Open Biol* **7:** 170073. doi:10.1098/rsob.170073

Hafez M, Hausner G. 2015. Convergent evolution of twintron-like configurations: one is never enough. *RNA Biol* **12:** 1275–1288. doi:10.1080/15476286.2015.1103427

Hallermayr A, Graf J, Koehler U, Laner A, Schönfeld B, Benet-Pagès A, Holinski-Feder E. 2018. Extending the critical regions for mutations in the non-coding gene *RNU4ATAC* in another patient with Roifman Syndrome. *Clin Case Rep* **6:** 2224–2228. doi:10.1002/ccr3.1830

He H, Liyanarachchi S, Akagi K, Nagy R, Li J, Dietrich RC, Li W, Sebastian N, Wen B, Xin B, et al. 2011. Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science* **332:** 238–240. doi:10.1126/science.1200587

Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. 2011. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7:** 497. doi:10.1038/msb.2011.28

Heremans J, Garcia-Perez JE, Turro E, Schlenner SM, Casteels I, Collin R, de Zegher F, Greene D, Humblet-Baron S, Lesage S, et al. 2018. Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. *J Allergy Clin Immunol* **142:** 630–646. doi:10.1016/j.jaci.2017.11.061

Horiuchi K, Perez-Cerezales S, Papasaikas P, Ramos-Ibeas P, López-Cardona AP, Laguna-Barraza R, Fonseca Balvís N, Pericuesta E, Fernández-González R, Planells B, et al. 2018. Impaired spermatogenesis, muscle, and erythrocyte function in U12 intron splicing-defective *Zrsr1* mutant mice. *Cell Rep* **23:** 143–155. doi:10.1016/j.celrep.2018.03.028

Jackson IJ. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res* **19:** 3795–3798. doi:10.1093/nar/19.14.3795

Levine A, Durbin R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res* **29:** 4006–4013. doi:10.1093/nar/29.19.4006

Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, Lin C-F, Makalowski W, Brown JWS, Jarmolowski A. 2004. Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell* **16:** 1340–1352. doi:10.1105/tpc.020743

Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, Thiruvahindrapuram B, Merico D, Jobling R, Nalpathamkalam T, et al. 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med* **20:** 435–443. doi:10.1038/gim.2017.119

Lopes-Ramos CM, Paulson JN, Chen C-Y, Kuijjer ML, Fagny M, Platig J, Sonawane AR, DeMeo DL, Quackenbush J, Glass K. 2017. Regulatory network changes between cell lines and their tissues of origin. *BMC Genomics* **18:** 723. doi:10.1186/s12864-017-4111-x

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27:** 1696–1697. doi:10.1093/bioinformatics/btr189

Madan V, Kanojia D, Li J, Okamoto R, Sato-Otsubo A, Kohlmann A, Sanada M, Grossmann V, Sundaresan J, Shiraishi Y, et al. 2015. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat Commun* **6:** 6042. doi:10.1038/ncomms7042

Markmiller S, Cloonan N, Lardelli RM, Doggett K, Keightley M-C, Boglev Y, Trotter AJ, Ng AY, Wilkins SJ, Verkade H, et al. 2014. Minor class splicing shapes the zebrafish transcriptome during development. *Proc Natl Acad Sci* **111:** 3062–3067. doi:10.1073/pnas.1305536111

Mazin P, Xiong J, Liu X, Yan Z, Zhang X, Li M, He L, Somel M, Yuan Y, Phoebe Chen Y-P, et al. 2013. Widespread splicing changes in human brain development and aging. *Mol Syst Biol* **9:** 633. doi:10.1038/msb.2012.67

Merico D, Roifman M, Braunschweig U, Yuen RKC, Alexandrova R, Bates A, Reid B, Nalpathamkalam T, Wang Z, Thiruvahindrapuram B, et al. 2015. Compound heterozygous mutations in the noncoding *RNU4ATAC* cause Roifman Syndrome by disrupting minor intron splicing. *Nat Commun* **6:** 8718. doi:10.1038/ncomms9718

Middleton R, Gao D, Thomas A, Singh B, Au A, Wong JJ-L, Bomane A, Cosson B, Eyras E, Rasko JEJ, et al. 2017. IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol* **18:** 51. doi:10.1186/s13059-017-1184-4

Niemela EH, Oghabian A, Staals RHJ, Greco D, Pruijn GJM, Frilander MJ. 2014. Global analysis of the nuclear processing of transcripts with unspliced U12-type introns by the exosome. *Nucleic Acids Res* **42:** 7358–7369. doi:10.1093/nar/gku391

Oegema R, Baillat D, Schot R, van Unen LM, Brooks A, Kia SK, Hoogeboom AJM, Xia Z, Li W, Cesaroni M, et al. 2017. Correction: human mutations in integrator complex subunits link transcriptome integrity to brain development. *PLoS Genet* **13:** e1006923. doi:10.1371/journal.pgen.1006923

Oghabian A, Greco D, Frilander MJ. 2018. IntEREst: intron-exon retention estimator. *BMC Bioinformatics* **19:** 130. doi:10.1186/s12859-018-2122-5

Padgett RA. 2012. New connections between splicing and human disease. *Trends Genet* **28:** 147–154. doi:10.1016/j.tig.2012.01.001

Putoux A, Alqahtani A, Pinson L, Paulussen ADC, Michel J, Besson A, Mazoyer S, Borg I, Nampoothiri S, Vasiljevic A, et al. 2016. Refining the phenotypical and mutational spectrum of Taybi-Linder syndrome. *Clin Genet* **90:** 550–555. doi:10.1111/cge.12781

Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, Valencia A, Tress ML. 2013. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res* **41:** D110–D117. doi:10.1093/nar/gks1058

Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F, Peterlongo P, Lacroix V. 2012. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* **13 (**Suppl 6**):** S5. doi:10.1186/1471-2105-13-S6-S5

Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, Meyer E, Duret L. 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol* **18:** 208. doi:10.1186/s13059-017-1344-6

Scamborova P, Wong A, Steitz JA. 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. doi:10.1128/MCB.24.5.1855-1869.2004

Schneider C, Will CL, Makarova OV, Makarov EM, Lührmann R. 2002. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol* **22:** 3219–3229. doi:10.1128/MCB.22.10.3219-3229.2002

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, 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. doi:10.1261/rna.053959.115

Scotti MM, Swanson MS. 2016. RNA mis-splicing in disease. *Nat Rev Genet* **17:** 19–32. doi:10.1038/nrg.2015.3

Shaheen R, Maddirevula S, Ewida N, Alsahli S, Abdel-Salam GMH, Zaki MS, Tala SA, Alhashem A, Softah A, Al-Owain M, et al. 2019. Genomic and phenotypic delineation of congenital microcephaly. *Genet Med* **21:** 545–552. doi:10.1038/s41436-018-0140-3

Shelihan I, Ehresmann S, Magnani C, Forzano F, Baldo C, Brunetti-Pierri N, Campeau PM. 2018. Lowry-Wood syndrome: further evidence of association with *RNU4ATAC*, and correlation between genotype and phenotype. *Hum Genet* **137:** 905–909. doi:10.1007/s00439-018-1950-8

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34:** 3955–3967. doi:10.1093/nar/gkl556

Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Hubbard TJ, Guigó R, Harrow J, Bertone P. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10:** 1177–1184. doi:10.1038/nmeth.2714

Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, Quesnel-Vallières M, Permanyer J, Sodaei R, Marquez Y, et al. 2017. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res* **27:** 1759–1768. doi:10.1101/gr.220962.117

Tarn WY, Steitz JA. 1996. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273:** 1824–1832. doi:10.1126/science.273.5283.1824

Turunen JJ, Will CL, Grote M, Lührmann R, Frilander MJ. 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. doi:10.1128/MCB.01928-07

Turunen JJ, Niemelä EH, Verma B, Frilander MJ. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA* **4:** 61–76. doi:10.1002/wrna.1141

Vanichkina DP, Schmitz U, Wong JJ-L, Rasko JEJ. 2018. Challenges in defining the role of intron retention in normal biology and disease. *Semin Cell Dev Biol* **75:** 40–49. doi:10.1016/j.semcdb.2017.07.030

Verma B, Akinyi MV, Norppa AJ, Frilander MJ. 2018. Minor spliceosome and disease. *Semin Cell Dev Biol* **79:** 103–112. doi:10.1016/j.semcdb.2017.09.036

Wang Y, Wu X, Du L, Zheng J, Deng S, Bi X, Chen Q, Xie H, Férec C, Cooper DN, et al. 2018. Identification of compound heterozygous variants in the noncoding *RNU4ATAC* gene in a Chinese family with two successive foetuses with severe microcephaly. *Hum Genomics* **12:** 3. doi:10.1186/s40246-018-0135-9

Wong JJ-L, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154:** 583–595. doi:10.1016/j.cell.2013.06.052

Wong JJ-L, Au AYM, Ritchie W, Rasko JEJ. 2016. Intron retention in mRNA: no longer nonsense: known and putative roles of intron retention in normal and disease biology. *Bioessays* **38:** 41–49. doi:10.1002/bies.201500117

Wu Q, Krainer AR. 1996. U1-mediated exon definition interactions between AT-AC and GT-AG introns. *Science* **274:** 1005–1008. doi:10.1126/science.274.5289.1005

Younis I, Dittmar K, Wang W, Foley SW, Berg MG, Hu KY, Wei Z, Wan L, Dreyfuss G. 2013. Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. *Elife* **2:** e00780. doi:10.7554/eLife.00780

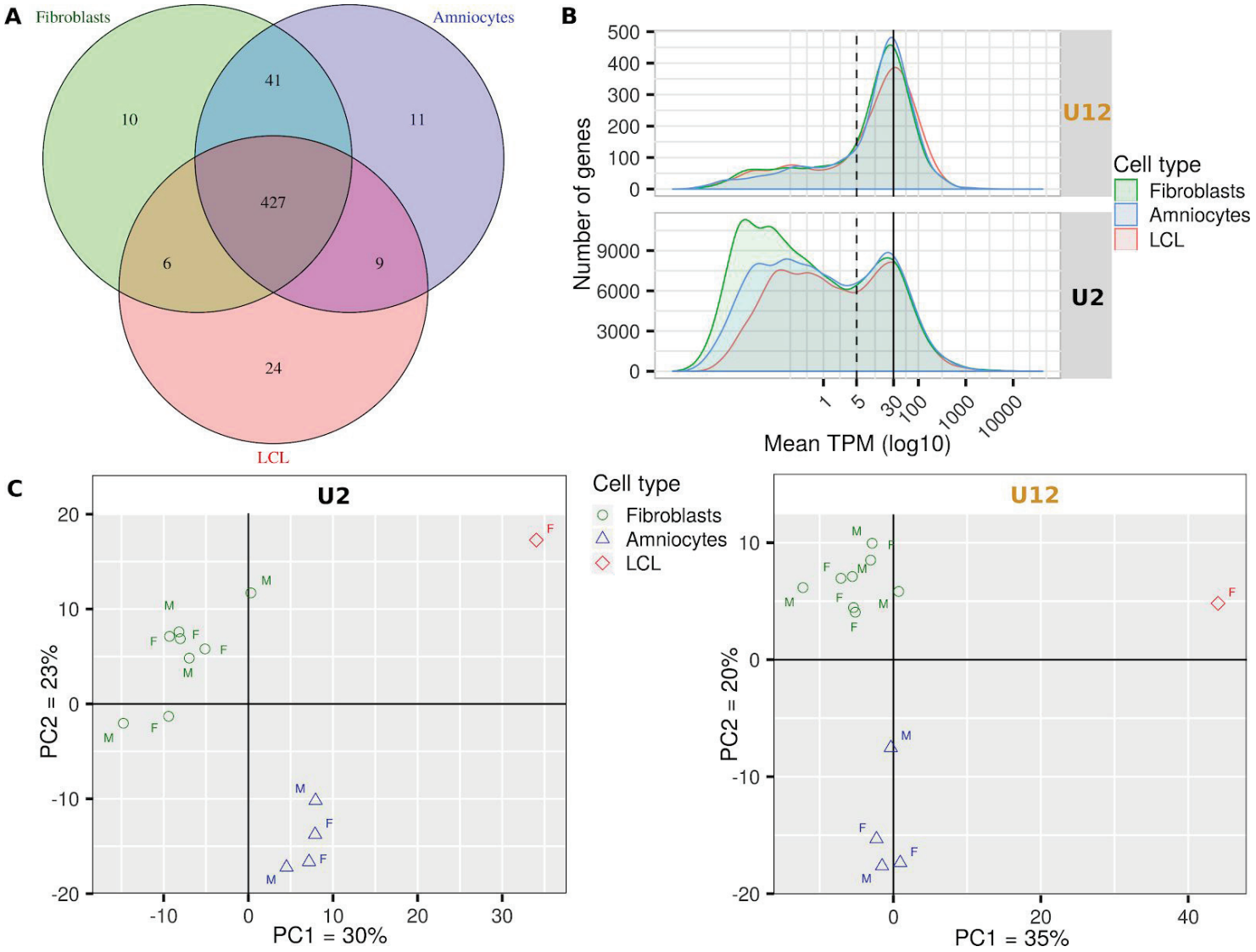

FIGURE S1 : Transcript levels of U12 and U2 genes in control fibroblasts, amniocytes and LCL. (*A*) Number of U12 genes expressed with a mean Transcript Per Million (TPM) value ≥ 5. (*B*) Distribution of non-zero mean TPM of U12 and U2 genes. The dashed and solid vertical lines indicates a TPM of 5 and 30 respectively. (*C*) Principal component analyses of U2 and U12 gene TPM values. Fibroblasts (8 samples), amniocytes (4 samples) and LCL (1 sample) were derived from tissues taken from control foetuses and children. The sex of the donor from which was derived each sample is indicated (M=Male, F=Female).

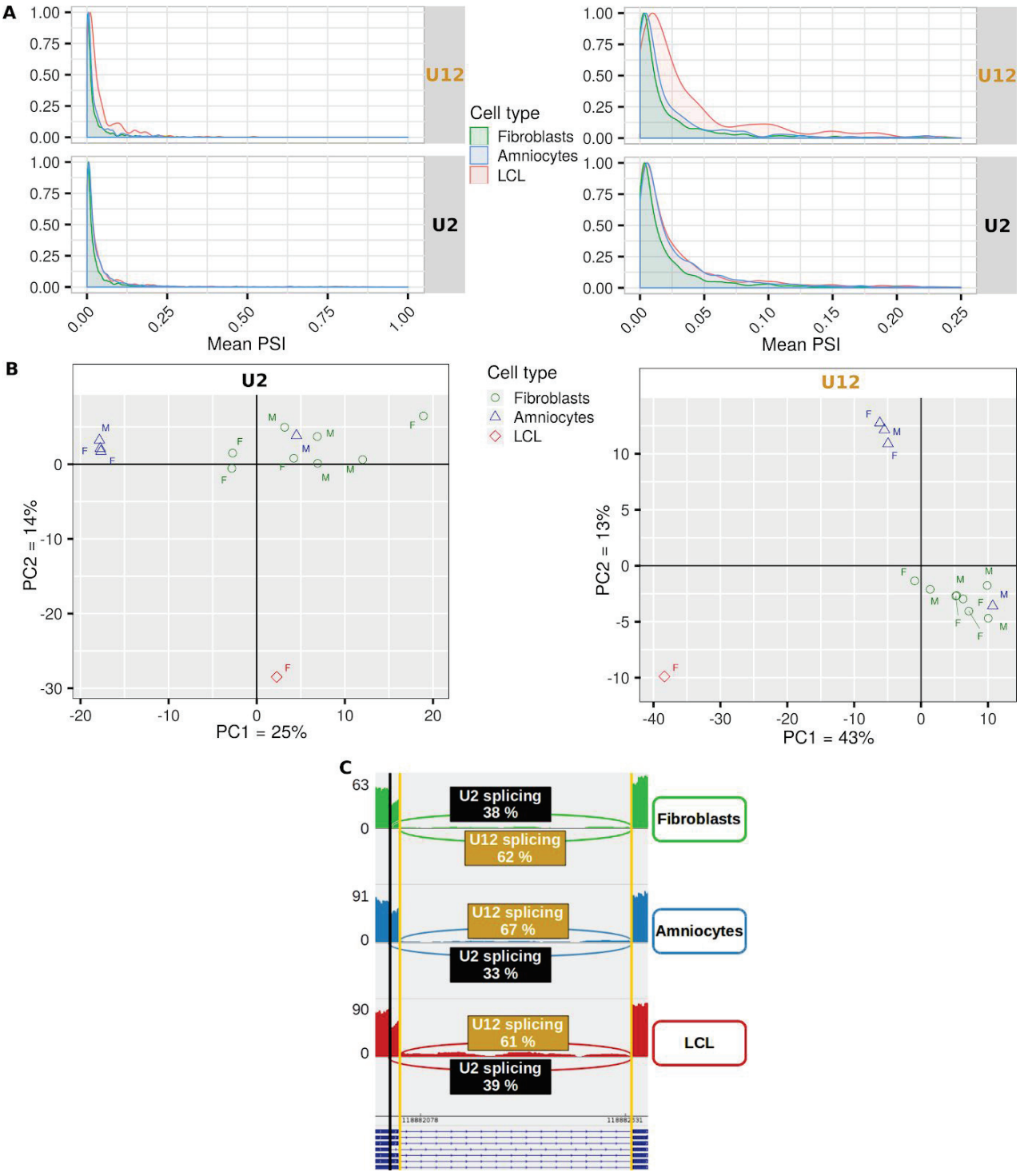

FIGURE S2 : Splicing of U12- and U2-type introns in control fibroblasts, amniocytes and LCL.
(*A*) Distribution of U12- and U2-type intron retention levels expressed with the Percent Spliced In metric (PSI) calculated for U12- (n=366) and U2-type (n=1887) introns in U12 genes (n=337) on a scale of 0 to 1 (left) or a zooming in from 0 to 0.25 (right). Poorly spliced introns (<5 reads covering the exon-exon junction on average for each cell-type) were filtered out. **(***B***)** Principal component analysis of U2- and U12-type intron mean PSI values. The same samples as in Fig. S1 were analysed. The sex of the donor from which was derived each sample is indicated (M=Male, F=Female). (*C*) Sashimi plots showing a U12/U2 splice site use switching event in the *CCDC84* gene. The y-axis corresponds to the mean coverage of each base of the genomic coordinates (x-axis). Reference annotations are given on the lowest part of the figure, with annotated exons and introns shown as thick and thin horizontal lines respectively. U12 and U2 splice sites are marked with yellow and black vertical bars respectively. Splice junction reads are drawn as arcs connecting a pair of exons. Mean percentage of reads supporting the splicing of either the U12- or U2-type intron are indicated in yellow and black boxes, respectively.

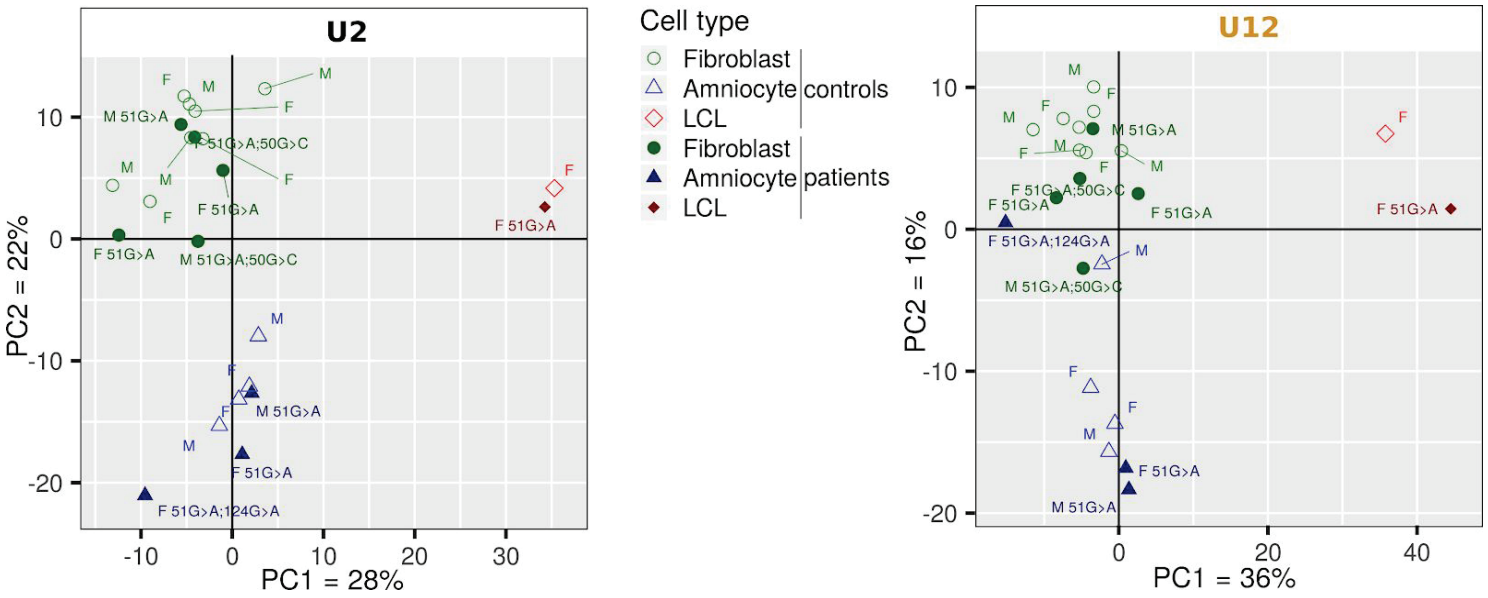

FIGURE S3 : Dominant patterns of U2 and U12 transcriptional gene expression in TALS patient and control cells. Principal component analyses of Transcripts Per Million values of U2 and U12 genes are presented. The same samples as in Fig. S1 were analysed for the control set, as well as fibroblasts (5 samples), amniocytes (3 samples) and LCL (1 sample) derived from tissues taken from TALS foetuses and children. The sex of the donor from which was derived each sample is indicated (M=Male, F=Female), as well as the *RNU4ATAC* mutation(s) for the patients' samples.

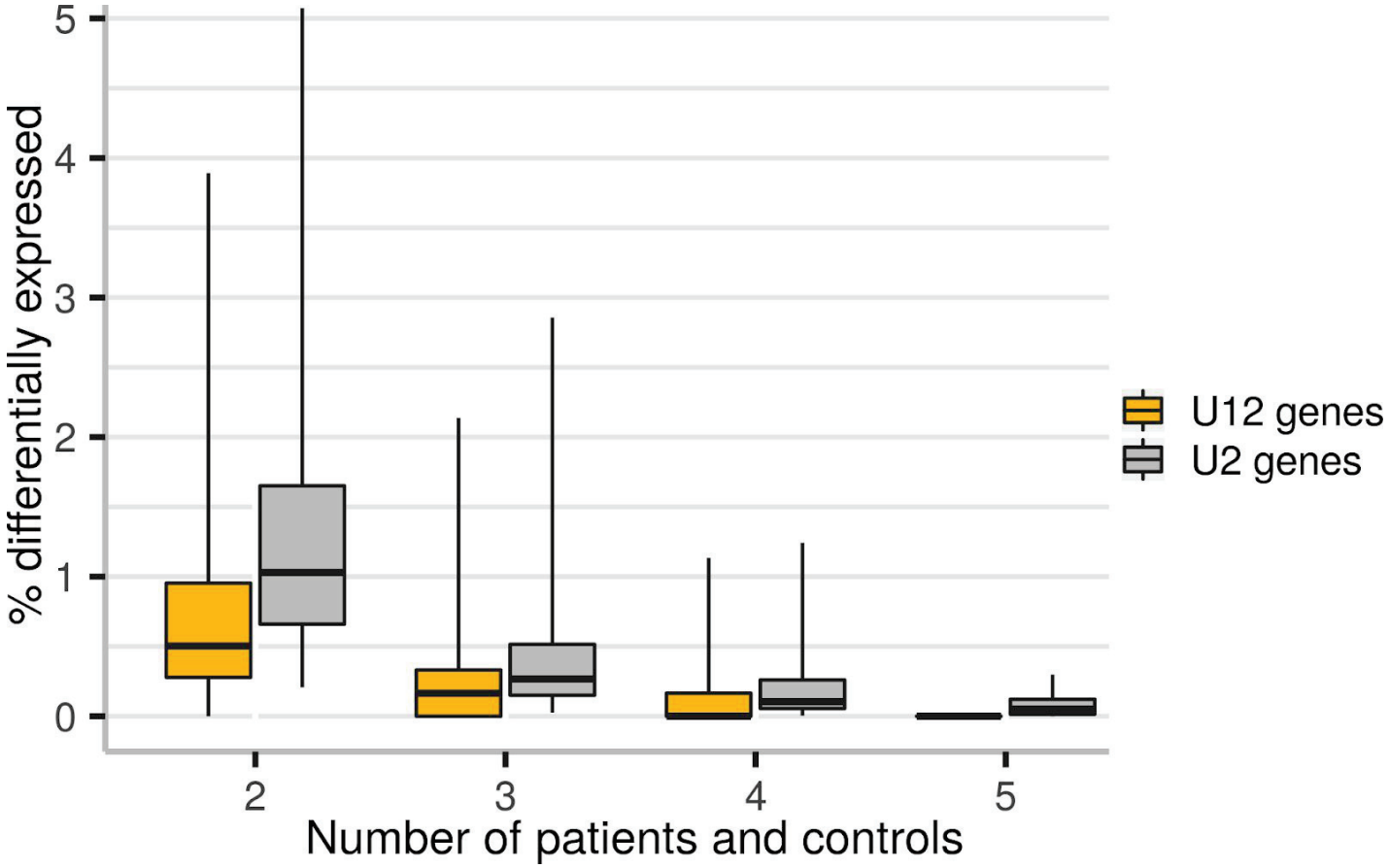

FIGURE S4 : Percentage of differentially expressed genes as a function of the number of patients' and controls' datasets analysed.

Boxplots of the percentage of differentially expressed genes identified by DESeq2 in the fibroblast datasets, using standard cutoffs (FDR $\leq$ 5%, $|\log2(FC)| \geq 2$). All possible combinations of TALS patients and controls respecting sex balance were tested. Two patients vs. two controls: 120 combinations; three patients vs. three controls: 220 combinations; four patients vs. four controls: 140 combinations; five patients vs. five controls: 24 combinations.

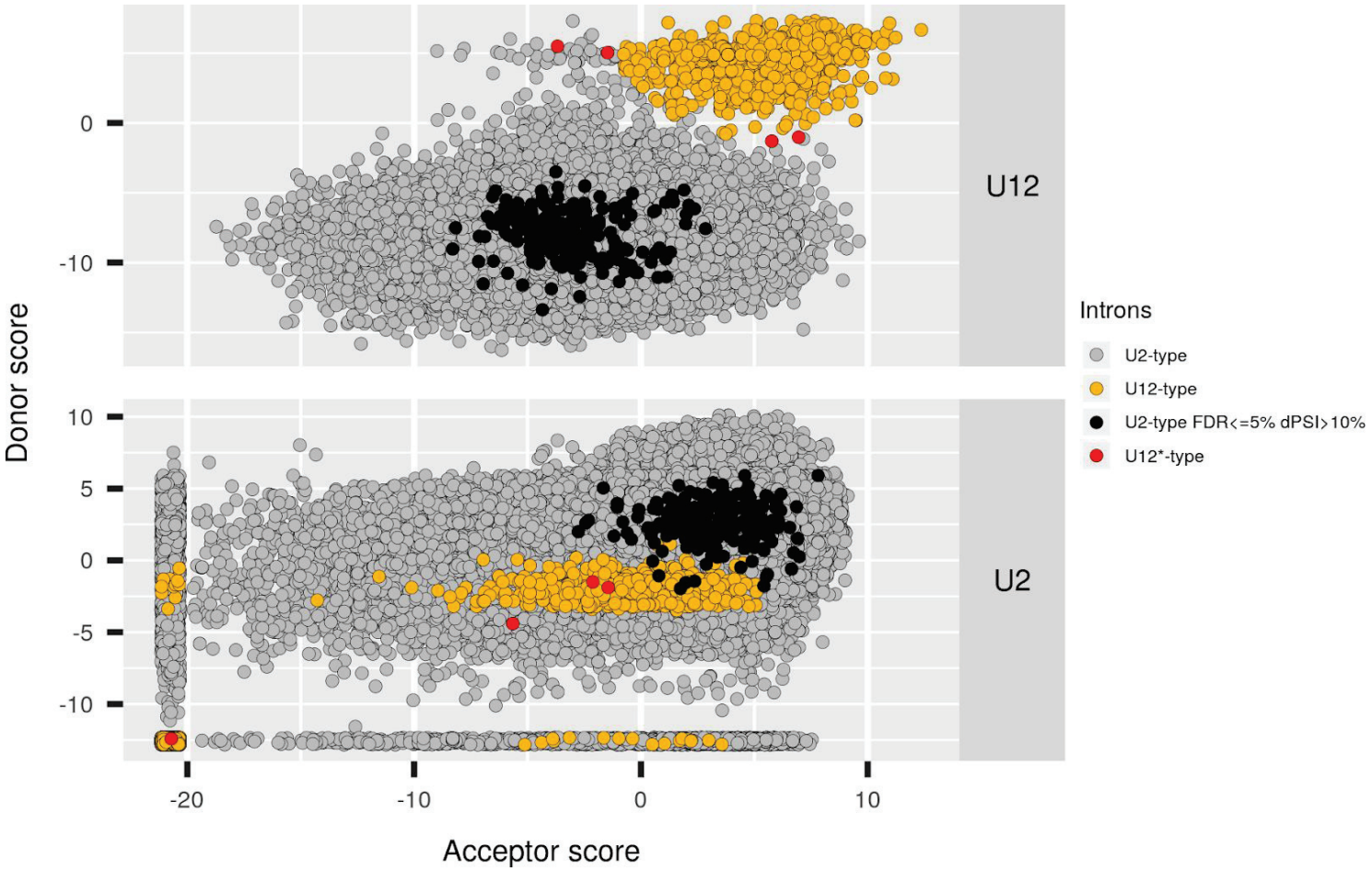

FIGURE S5 : Splice sites scores for all U12- and U2-type introns.

Each annotated intron is plotted with respect to its U12 (top) or U2 (bottom) splice site scores, as calculated with T. Alioto's present scripts. U12*-type introns: U2-type introns proposed to be reclassified as U12-type introns, namely: *RECQL5, DERL2, KIAA0556* and *LZTR1*. For a better visualisation, introns without computed U2 score for the donor and/or acceptor splice site(s) (sequence(s) too divergent from the consensus, score = -100 by default) were given a new score (with a jittering effect) of -12.58 and -20.74 respectively, corresponding to the minimum observable score for the splice site type (donor and acceptor) minus 1.

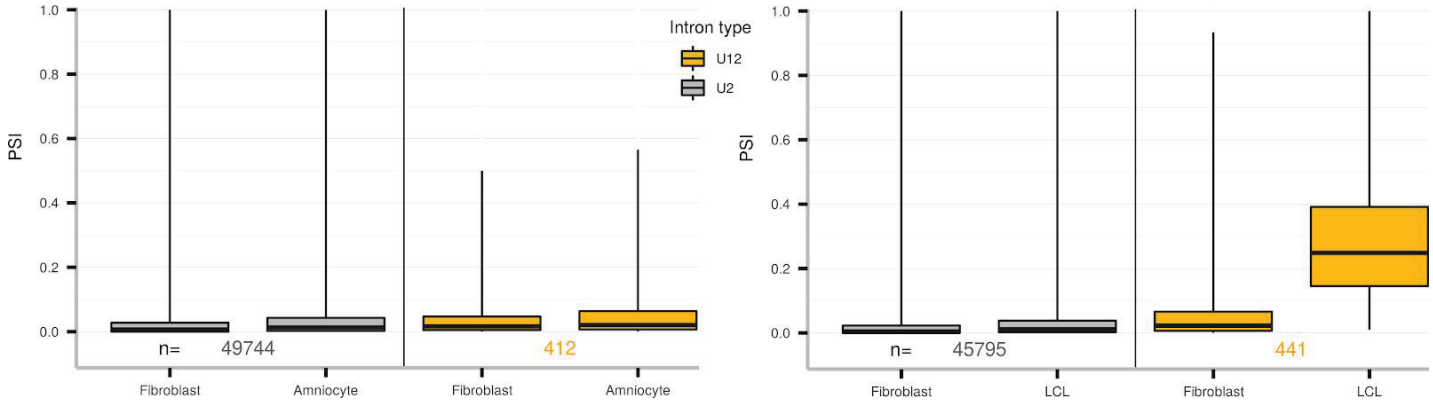

FIGURE S6 : Comparison of U2- and U12-type intron retention levels in two different cell-types derived from the same TALS patient.

Boxplots of U2- and U12-type intron PSI values (PSI-boxplots) of datasets obtained from two cell-types derived from the same TALS patient: fibroblasts and amniocytes (left, TALS2), or fibroblasts and LCL (right, TALS6). The two patients, described in Table 1, are both homozygous carriers of g.51G>A. The number of U2- and U12-type introns analysed (i.e. with a sufficient coverage in each sample) are indicated.

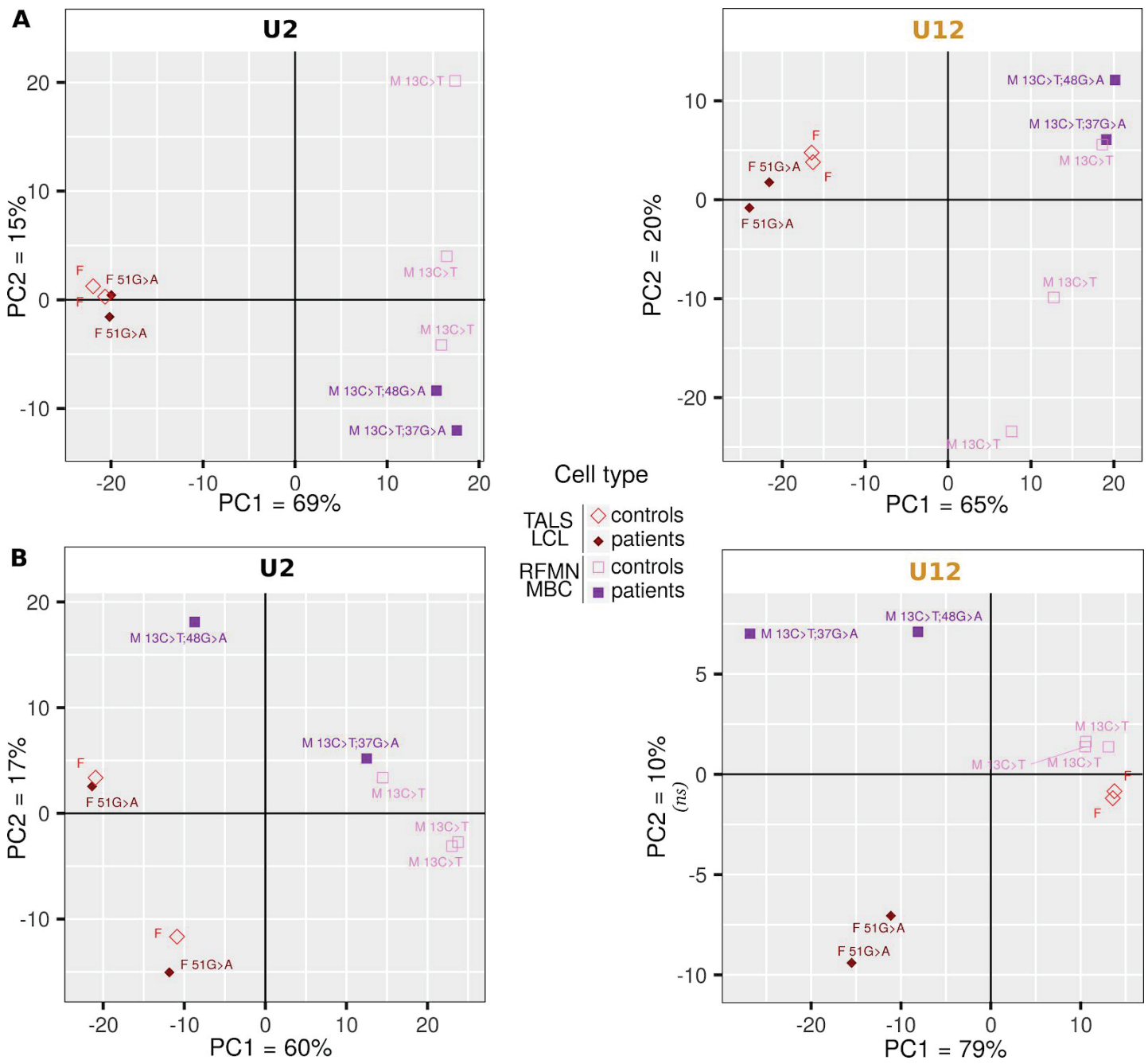

FIGURE S7 : Dominant patterns of gene expression and intron retention in TALS and RFMN patient and control blood cells.

(*A*) Principal component analyses of TPM values of U2 and U12 genes, and (*B*) PCA of mean PSI values of U2- and U12-type introns are presented. The datasets analysed are the following: 1 TALS patient's and 1 control's LCL datasets (two technical replicates for each) and 2 RFMN patients' and 3 related controls' MBC datasets. The sex of the donor from which was derived each sample is indicated (M=Male, F=Female), as well as the *RNU4ATAC* mutation(s) for the patients' samples. TPM: Transcript Per Million; LCL: lymphoblastoid cell line; MBC: mononuclear blood cells; *ns*: not significant (the explained variance of the axis is smaller or equal to the explained variance of our randomised data, see Methods). The same patterns were obtained when the analyses were conducted with the LCL datasets of the extended study and those obtained for unrelated individuals of the RFMN collection.

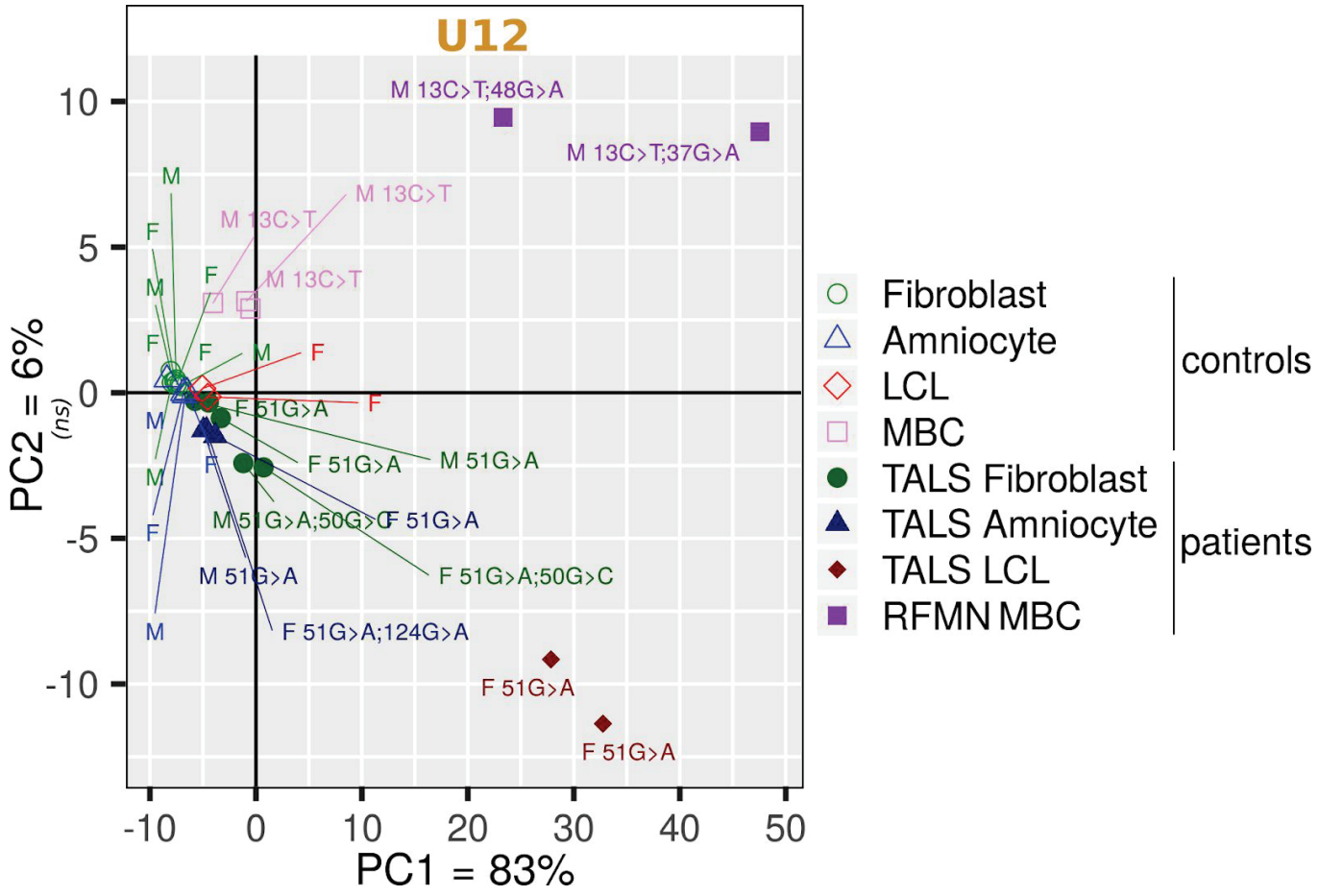

FIGURE S8 : Patterns of U12-type intron retention in TALS or RFMN patient and control cells.

PCA of the most variable mean PSI values of U12-type introns are presented. The datasets analysed are the following: 5 TALS patient and 8 control fibroblast datasets, 3 TALS patient and 4 control amniocyte datasets, 1 TALS patient and 1 control LCL datasets (two technical replicates for each) and 2 RFMN patient and 3 related control MBC datasets. The sex of the donor from which was derived each sample is indicated (M=Male, F=Female), as well as the *RNU4ATAC* mutation(s) for the patients' samples. LCL: lymphoblastoid cell line; MBC: mononuclear blood cells; *ns*: not significant (the explained variance of the axis is smaller or equal to the explained variance of our randomised data, see Methods).

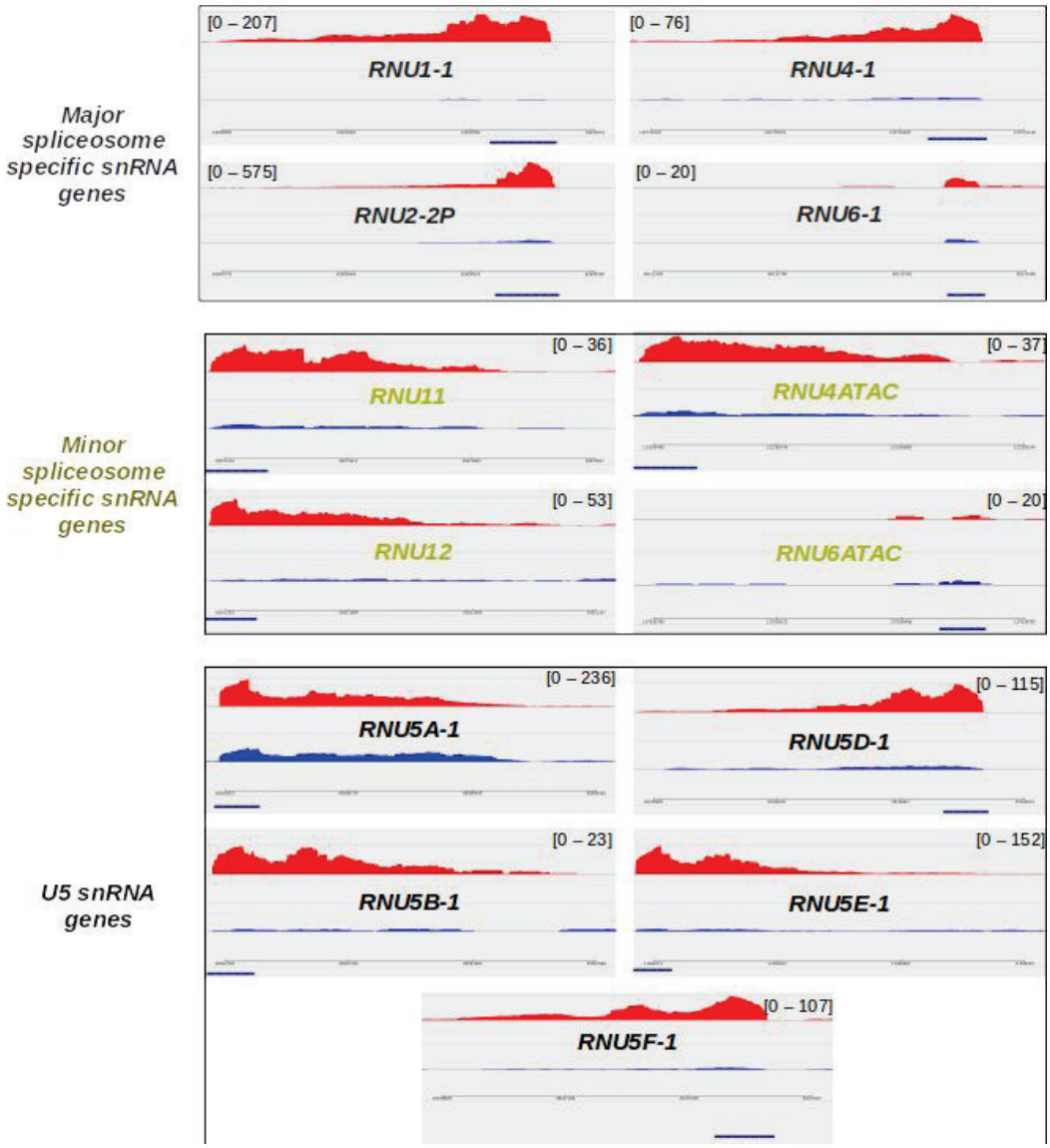

FIGURE S9 : Genomic read coverage along minor and major spliceosome snRNA gene regions in control and TALS LCL samples.

The read coverage from the LCL patient (red) and control (blue) datasets over each spliceosomal snRNA gene region is shown. The location of each annotated snRNA gene is indicated by a thick blue line along the genome position. The read coverage scale across the genomic window is indicated for the TALS LCL sample at the top left or right corner of each panel. The multiple *RNU1*, *RNU2*, *RNU4* and *RNU6* gene copies, organised as tandem arrays, are shown at a unique location, while the multiple *RNU5* loci are shown (*RNU5A/5B/D/E/F*). *RNU2-2P* corresponds to the *RNU2-1* gene in the ensembl75 version of the annotation.

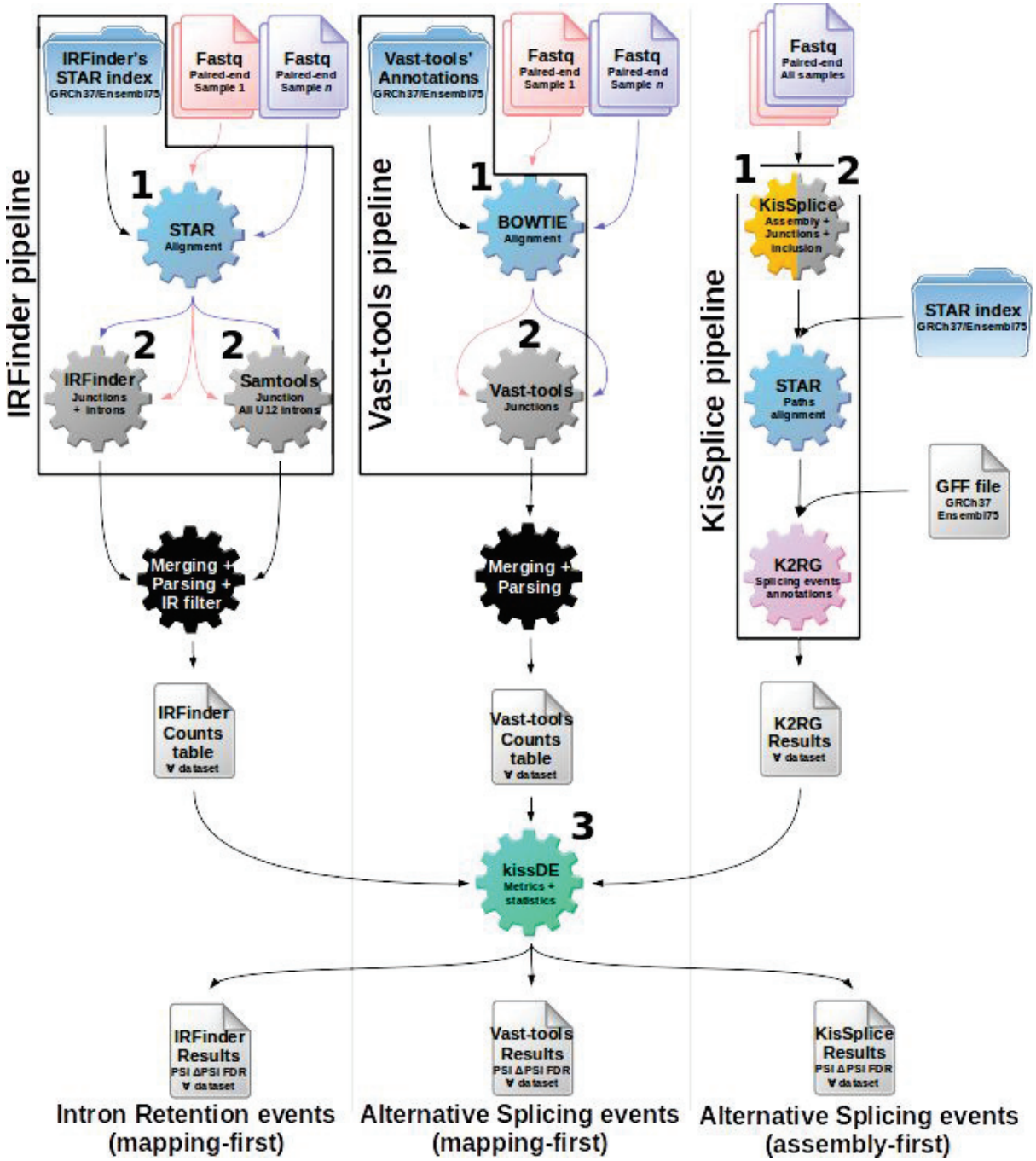

FIGURE S10 : Bioinformatics analysis overview.

Workflow of the three pipelines used for the splicing analysis. Input read files are modelled with a red or blue file icon (fastq files). Each step underwent by a sample is modelled by a red or blue arrow, corresponding to its fastq files, or by a black arrow for a step underwent by all samples at once. Each gear represents a published tool or in-house python script : reads alignment, reads quantification, reads assembly, statistical analysis, counts formatting (in-house scripts) and splicing events annotation are represented by blue, grey, yellow, green, black and red gears, respectively. Numbers indicate which main goal is achieved by the tool : 1 = read alignment/assembly; 2 = read quantification on exon-exon/intron-exon junctions (noted as "Junctions") and/or on the included part of the event (noted as "Introns" for IR and "Inclusion" for other alternative splicing events); 3 = PSI/ΔPSI/FDR computation. K2RG = KisSplice2RefGenome.

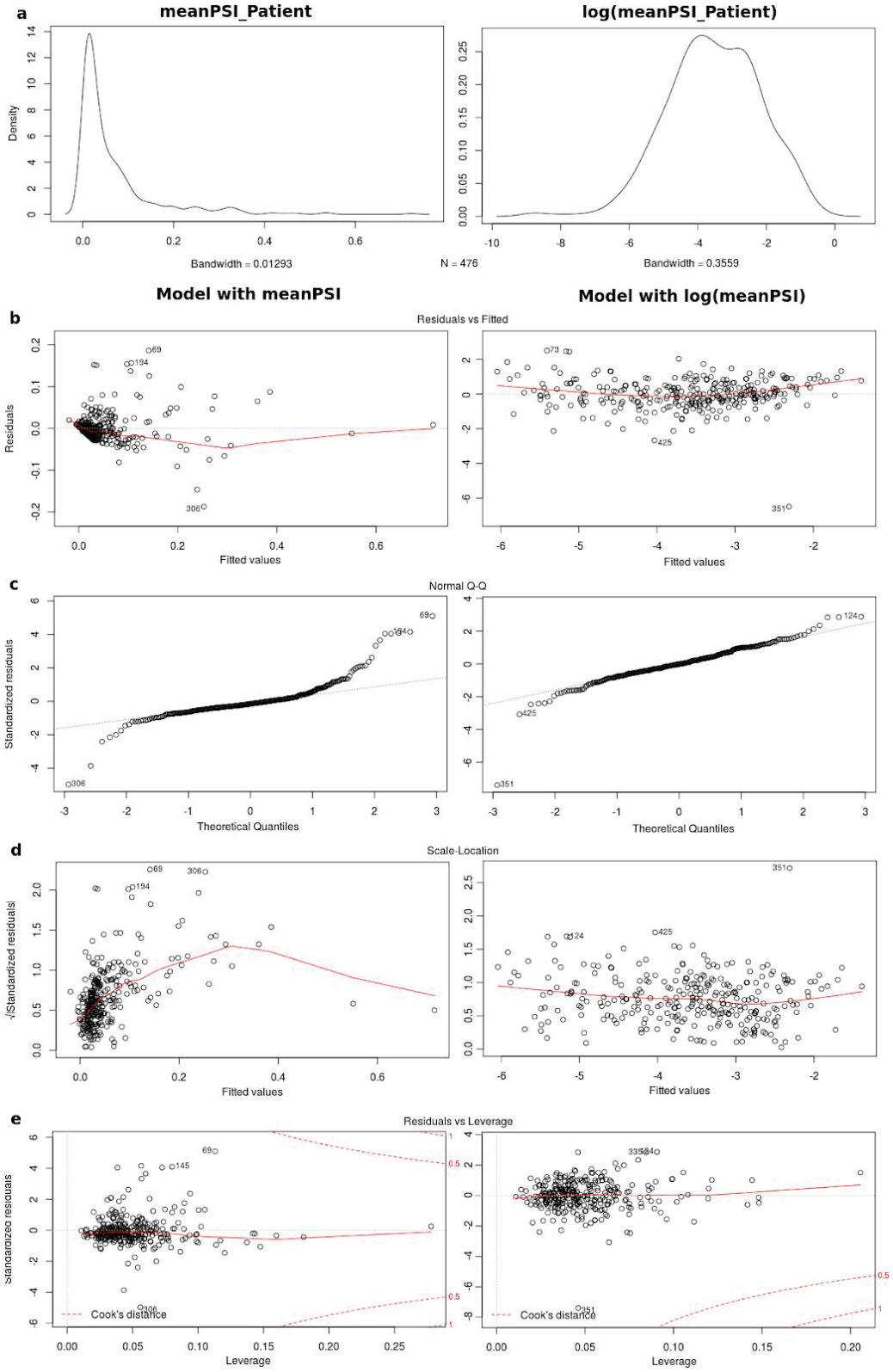

FIGURE S11 : Linear models' diagnostic plots.

(*A*) Plots of the patient U12-type intron mean PSI (left) and log(mean PSI) (right) distribution. (*B*), (*C*), (*D*), (*E*) Diagnostic plots (plot(lm(model)) of the complete model explaining the U12-type intron patient mean PSI, using either the patient and controls U12-type intron mean PSI (left) or log(mean PSI) (right). (*B*) Residual vs. Fitted plot: equally spread residuals around an horizontal line indicate linear relationship between the response variable and predictors. (*C*) Normal Q-Q plot: residuals following a straight line indicate normal distribution. (*D*) Scale-Location plot: equally spread residuals around an horizontal line indicate homoscedasticity. (*E*) Residuals vs. Leverage plot: identify possibly influential outliers observations (outside of a dashed red line).

C. Discussion

Cette publication apporte des informations importantes sur nos connaissances de l'épissage mineur autant en condition physiologique que pathologique. Physiologiquement, les introns U12 semblent aussi bien épissés que les U2 dans des cellules d'enfant ou de fœtus. Nous observons beaucoup d'événements d'épissage U12 (deux introns U12 alternativement épissés) impliquant un changement de site accepteur uniquement, et beaucoup d'événements d'épissage U12/U2 (un intron U12 ou U2 alternativement épissé) impliquant au moins un changement de site donneur. Ces résultats sont compatibles avec le fait que le site donneur U12 est très conservé et différent du site U2, contrairement au site accepteur, qui pourrait ainsi être reconnu par les deux spliceosomes.

Sans surprise, nous trouvons que l'épissage mineur est globalement affecté chez les patients TALS comparés aux contrôles : la qualité de l'épissage des introns U12 est réduite. De plus, il semblerait que la rétention d'un intron U12 puisse induire, dans certains cas, la rétention d'un des deux introns U2 adjacents (apparemment sans privilégier celui situé en amont ou en aval). Cette observation est un argument supplémentaire supportant l'hypothèse que les spliceosomes mineur et majeur interagissent ensembles lors de l'épissage (Horiuchi et al., 2018; Lewandowska et al., 2004; Tapial et al., 2017; Wu and Krainer, 1996). Il arrive aussi que l'intron U12 ne soit pas retenu dans le transcrit mature, puisqu'un intron U2 le chevauchant est épissé à la place. C'est notamment le cas dans le gène *CCDC84*, de fonction inconnue, pour lequel un site donneur U2 est très largement privilégié chez les patients comparés aux contrôles (Figure 26). Il s'agit de l'événement d'épissage retrouvé chez l'ensemble des patients TALS (tous type-cellulaires confondus) qui a l'effet le plus grand.

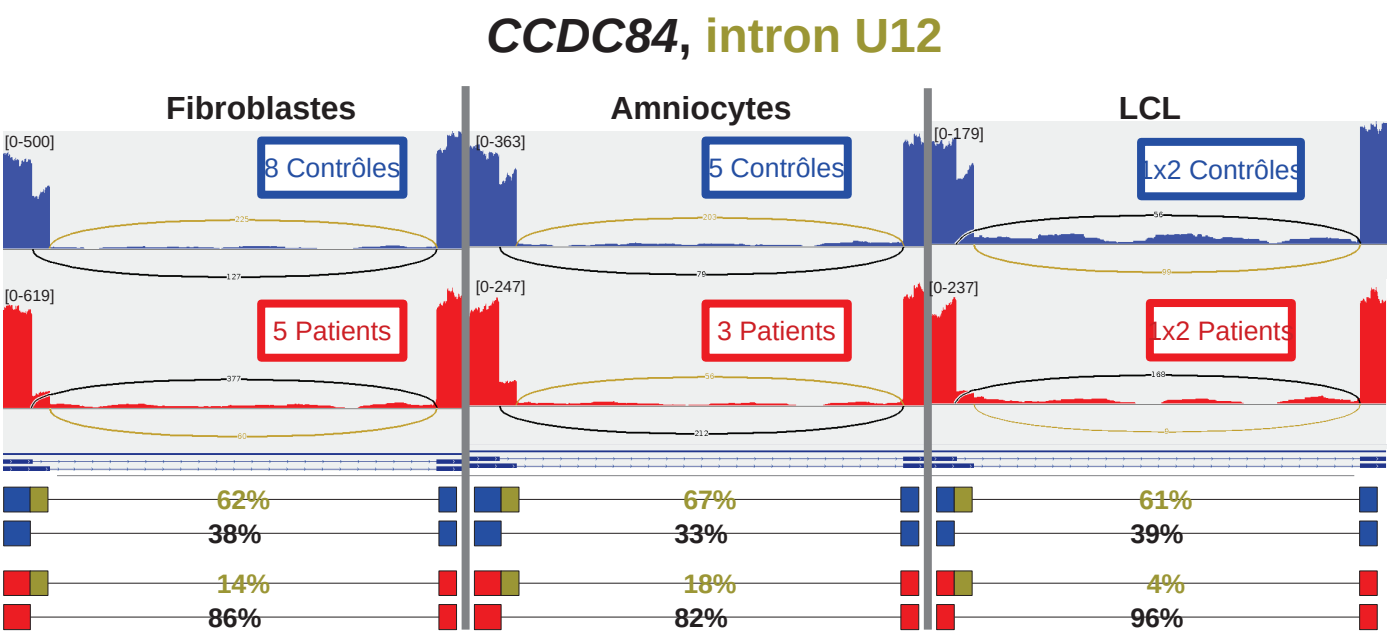

Figure 26 : Utilisation préférentielle d'un site d'épissage U2 chez les patients TALS.
Donneur alternatif U12/U2 dans le gène *CCDC84*, chez les patients TALS (rouge) et les contrôles (bleu). Les deux transcrits possibles, épissant soit l'intron U12 (jaune) ou l'intron U2 chevauchant (noir) sont représentés, avec leur abondance relative, sous chaque sashimi plot et pour chaque condition.

L'aspect tissu-dépendant des conséquences moléculaires du TALS est surprenant tant le déficit en épissage mineur est amplifié dans le LCL comparé aux fibroblastes et amniocytes (Figure 27). De nombreux événements d'épissages autres que des rétentions d'intron sont fréquemment observés dans les LCL, certains utilisant des sites d'épissages U2 *de novo* situés à proximité des sites d'épissage U12 (Figure 27, traits noirs), un phénomène que nous avons nommé « splice site switching ». La différence entre les types cellulaires de patients TALS se traduit par une forte augmentation du PSI des patients pour les rétentions d'intron U12. En effet, alors que seulement 11 et 14 % des introns U12 sont significativement et fortement retenus (10 % de différence entre le PSI patient et contrôle) dans les données RNA-seq de fibroblastes et d'amniocytes respectivement, ce pourcentage atteint 77 % dans le jeu de données des LCL. Cette spécificité est d'autant plus intrigante que le profil d'épissage de notre patient TALS LCL est quasiment identique à celui de patients RFMN, malgré la multitude de facteurs qui pourrait les différencier. Bien que les signes cliniques de patients atteints du syndrome de Taybi-Linder et du syndrome de Roifman soient très différents, leur phénotype moléculaire pourrait-il être identique lorsque les mêmes types cellulaires sont comparés ? Existe-t-il des tissus dans lesquels le profil transcriptomique des patients TALS et RFMN est très différent ? Seules des LCL d'autres patients TALS ou des

fibroblastes/amniocytes de patients RFMN permettraient de confirmer cette tissu-spécificité et de déterminer les caractéristiques moléculaires de l'un et l'autre des syndromes.

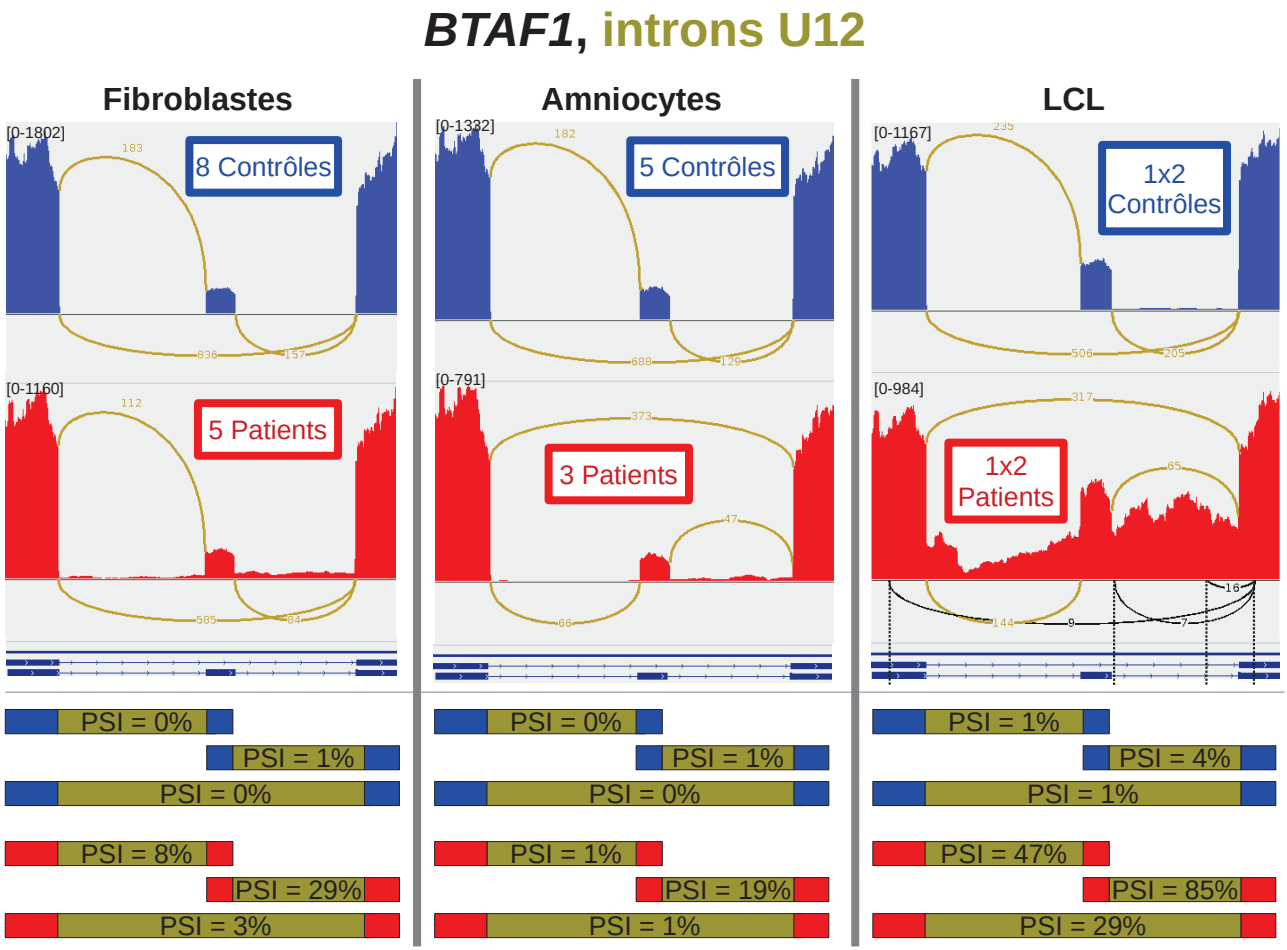

Figure 27 : Tissu-spécificité des rétentions d'intron U12 et de l'utilisation de sites U2 *de novo*. Sashimis plots (haut) présentant les reads provenant de fibroblastes (gauche), amniocytes (milieu) et LCL (droite) d'individus contrôles (bleu) et patients TALS (rouge), pour les trois introns U12 situés dans le gène *BTAF1*. Les jonctions reliant des sites d'épissage mineur et majeur sont colorées en jaune et noir, respectivement. Un trait vertical pointillé noir indique la position d'un site d'épissage U2 *de novo* sur l'annotation. La partie basse du graphique indique les pourcentages de rétention des trois introns U12 chez les contrôles (bleu) et chez les patients (rouge).

Quel pourrait être le devenir des transcrits contenant une rétention d'intron U12 ? Ils pourraient être retenus dans le noyau (detained introns), ou être dégradés par l'exosome nucléaire ou cytoplasmique. Cependant, nous observons des rétentions d'introns mais pas de différentiel d'expression. Cette expression est mesurée à partir de l'extraction de tous les ARN nucléaires et cytoplasmiques disposant d'une queue polyA, il semble donc peu probable que les transcrits matures contenant une rétention d'intron U12 soient dégradés, que ce soit dans le noyau ou dans le cytoplasme. Les transcrits aberrants semblent donc capables d'échapper aux

contrôles qualités nucléaires et cytoplasmique, notamment le NMD (il est d'ailleurs intéressant de noter que *UPF1*, protéine centrale du NMD, possède un intron U12).

Pour quantifier la fraction d'ARN aberrant produite par un gène et la fraction d'ARN aberrant dégradée par des contrôles qualités cytoplasmiques et/ou nucléaire des ARN, nous avons mené une expérience de fractionnement permettant de séparer les fractions nucléaires et cytoplasmiques. Ceci nous a permis d'étudier la quantité d'intron retenue dans chaque fraction. Nos résultats préliminaires indiquent que le contrôle qualité cytoplasmique dégrade le même pourcentage de transcrits aberrants entre patients et contrôles. La quantité de transcrits contenant une rétention d'intron U12 est augmentée dans le noyau de cellules de patients TALS, ce qui pourrait être le résultat d'une plus grande production de transcrits aberrants, ou bien d'une plus grande proportion de detained intron, ou des deux.

Bien que certains transcrits contenant une rétention d'intron U12 échappent aux contrôles qualités nucléaires et cytoplasmiques, par un procédé non caractérisé, il est tout de même improbable qu'ils puissent coder une protéine fonctionnelle. Ainsi, bien que le niveau de transcrits ne varie pas entre patients TALS et contrôles, la quantité de protéines fonctionnelles est vraisemblablement diminuée d'un niveau équivalent à la proportion de transcrits aberrants trouvé en plus dans les cellules des patients, comparé aux cellules des contrôles.

Concernant l'étude d'enrichissement sur les gènes U12 sujets à de fortes rétentions d'introns U12, nous avons été étonnés de trouver plusieurs processus biologiques impliqués dans l'immunité, puisque qu'aucun déficit immunitaire n'a été décrit chez les TALS, alors qu'il est un des signes caractéristiques des RFMN. Des études sur de nouveaux patients TALS, reprenant le protocole d'analyse utilisé pour caractériser le profil immunitaire de patients RFMN, indiquent qu'un défaut immunitaire existe bel et bien chez les TALS.

Le profil d'épissage étant très particulier et complexe lors de la neurogenèse (Raj and Blencowe, 2015), et le phénotype moléculaire apparaissant comme extrêmement dépendant du type-cellulaire, nos résultats ne nous incitent pas à nous prononcer sur une quelconque chaîne de causalité moléculaire permettant de lier l'épissage mineur au développement embryonnaire. Pour aller plus loin, des expériences sur des cellules souches pluripotentes induites ou sur un modèle animal devront être réalisées. En revanche, ce travail consolide l'hypothèse que les spliceosomes majeur et mineur interagissent lors de l'épissage d'introns U2 et U12 proches, et prodigue de nouveaux exemples d'utilisation de sites d'épissage majeurs à proximité de sites mineurs dans le contexte d'un spliceosome mineur défaillant. De

plus, pour la première fois à notre connaissance, des exemples physiologiques d'alternance entre l'utilisation de sites d'épissage U2 et U12 ont été caractérisés, laissant supposer qu'il existe naturellement une compétition entre spliceosome majeur et mineur pour épisser certains introns. Enfin, nous montrons que les défauts d'épissage mineur dans le TALS sont très fortement dépendant du tissu étudié.

4 Discussion et perspectives

L'objectif de cette thèse était de caractériser précisément le profil transcriptomique de cellules de patients atteints du Syndrome de Taybi-Linder (TALS), maladie du développement embryonnaire dans laquelle le spliceosome mineur est défectueux. Pour cela, j'ai mis en place un pipeline bioinformatique permettant, à partir de données de séquençage de 2ème génération (reads courts), de réaliser une analyse différentielle de l'expression des gènes et des événements d'épissage alternatifs entre patients et contrôles, avec une attention particulière donnée aux rétentions d'intron (IR). Par rapport à d'autres études transcriptomiques sur des maladies liées à une déficience de l'épissage mineur, nous avons l'originalité d'utiliser des méthodes d'analyse mapping-first et une méthode assembly-first, nous permettant de caractériser plus finement les défauts d'épissage. Nous avons aussi l'avantage de posséder le plus grande cohorte internationale de patients pour des mutations sur *RNU4ATAC*, renforçant la puissance de notre analyse statistique.

A. Analyse des événements d'épissage

a. KisSplice :assembleur local

Déterminer l'identité et l'abondance des transcrits à partir de reads courts est une tâche bioinformatique toujours non-résolue à ce jour. La principale difficulté est que la faible taille des reads ne leur permet pas, le plus souvent, de s'aligner sur plus d'une jonction exon-exon. Donc, même si il est possible de trouver l'ensemble de ces jonctions, il est souvent très difficile de phaser les exons, c'est-à-dire de déterminer quels exons proviennent d'un même transcrit, et encore plus difficile de quantifier les différents transcrits (The RGASP Consortium et al., 2013b). C'est pourquoi nous avons privilégié une méthode locale, KisSplice (Sacomoto et al., 2012), qui tente uniquement de trouver les événements d'épissage sans émettre de prédiction sur la composition en exons des transcrits complets, contrairement à des méthodes plus populaires comme OASES (Schulz et al., 2012), Trinity (Grabherr et al., 2011) ou Cufflinks (Trapnell et al., 2012). KisSplice est hébergé au Laboratoire de Biométrie et Biologie Évolutive, où j'ai effectué une partie de ma thèse, et j'ai ainsi pu participer à son développement. Durant les prochaines années, je devrais poursuivre mon implication dans le développement de KisSplice et des outils qui y sont associés : KisSplice2RefGenome et kissDE. Nous envisageons également d'améliorer la visibilité du pipeline KisSplice en le

rendant disponible sur le cloud de l'Institut Français de Bioinformatique et de caractériser plus globalement les cas d'utilisation où KisSplice s'avère supérieure aux autres méthodes d'analyse bioinformatique de l'épissage. En effet, les travaux que nous avons publiés (Benoit-Pilven et al., 2018) étaient restreints à l'analyse de sauts d'exons. Or, nous avons mis en évidence au cours de cette thèse que de nouveaux sites donneurs et accepteurs alternatifs non-annotés peuvent aussi être découverts par KisSplice. Bien que des méthodes mapping-first récentes comme LeafCutter (Li et al., 2016) ou MAJIQ (Vaquero-Garcia et al., 2016) permettent elles aussi de trouver de nouveaux sites d'épissage sans utiliser d'annotation de référence, il semble que l'approche assembly-first soit plus sensible, notamment pour les changements impliquant des petits segments génomiques. Une comparaison de notre méthodes assembly-first avec ces méthodes mapping-first paraît donc nécessaire pour identifier dans quelles situations KisSplice est la seule méthode à identifier des sites d'épissage. Ce travail est actuellement en cours de réalisation.

b. Séquençage de 3ème génération

En plus d'utiliser des méthodes d'analyses locales de l'épissage, plusieurs alternatives existent pour résoudre le problème de reconstruction des transcrits. L'approche la plus prometteuse est le séquençage de 3ème génération. Les technologies les plus populaires sont actuellement celles de Pacific Bioscience (PacBio) (Eid et al., 2009) et d'Oxford Nanopore Technologies (ONT) (Branton et al., 2009). Les reads produits par ces technologies peuvent théoriquement couvrir l'intégralité des transcrits, rendant leur identification aisée, sans pour autant impacter significativement le coût de séquençage. De plus, il est possible, avec la technologie ONT, de séquencer directement l'ARN (sans passer par une étape de rétrotranscription en ADNc), ce qui permet notamment d'améliorer la précision de la quantification des transcrits (Sessegolo et al., 2019). Cependant, ces technologies souffrent d'une faible précision de séquençage (~10% d'erreurs), rendant l'identification exacte des sites d'épissage difficile. Une autre limitation est la faible profondeur de séquençage (comparée aux méthodes de 2ème génération), qui peut empêcher la détection de transcrits rares. De nouvelles méthodes dites hybrides (utilisant des reads longs et courts) sont actuellement privilégiées.

Dans ce contexte de l'analyse de longs reads ou de données hybrides, des développements bioinformatiques sont encore nécessaires. C'est pour répondre à ce manque qu'est né en 2016 le projet Algorithmes et outils logiciels pour le Séquençage d'ARN de Troisième génERation (ASTER), qui se focalise sur l'analyse de données transcriptomiques provenant d'ONT. J'ai

eu l'occasion, au court de ma thèse, de participer aux réunions du projet ASTER organisées à Lille, Evry et Lyon. Ma contribution dans ce projet a pour l'instant été l'écriture d'un script permettant de convertir un alignement sur le transcriptome en un alignement sur le génome, qui permet, dans le contexte de l'alignement de reads longs, de caractériser précisément les sites d'épissages utilisés. Le développement d'outils bioinformatiques pour l'analyse de données de séquençage de 3ème génération est d'autant plus difficile que les technologies évoluent extrêmement vite ; toutefois, ces méthodes de séquençage sont très prometteuses, le taux d'erreur diminuant d'année en année.

B. Profil transcriptomique des patients TALS

Pour analyser l'épissage à partir de reads courts, nous avons utilisé deux approches complémentaires (mapping- et assembly-first) et nous nous sommes focalisés sur la détection de rétentions d'intron (IR), conséquence attendue du défaut d'épissage mineur. Comme les IR sont particulièrement difficiles à détecter et à quantifier, une méthode récente, IRFinder (Middleton et al., 2017), a été utilisée. KissDE (Clara Benoit-Pilven, 2018), un outil d'analyse statistique au développement duquel j'ai participé, est utilisé pour identifier les épissages différentiels entre nos deux conditions, les patients et les contrôles. KissDE n'est pas restreint à l'analyse des sorties de KisSplice et peut être appliqué en sortie de toute méthode permettant de quantifier des paires de variants.

a. Qualité de l'épissage des introns U12

Nous trouvons que les introns U12 sont plus retenus chez les patients TALS que chez leurs contrôles associés, mais à des niveaux extrêmement différents en fonction du tissu observé (Figure 28). Les rétentions d'introns U12 dans les fibroblastes et amniocytes sont bien visibles, mais d'une ampleur très modérée par rapport aux rétentions d'intron observées dans les cellules de lignée lymphoblastoïde (LCL) ou les cellules sanguines mononucléées d'une autre pathologie associée à une déficience du spliceosome mineur, le Syndrome de Roifman (RFMN). Bien que devant être confirmée, cette apparente tissu-spécificité doit être prise en compte pour déterminer les processus physiopathologiques du TALS, et démontre l'importance d'étudier divers tissus à divers stades de développement dans le cadre d'études sur l'épissage, afin de ne pas généraliser à tout un organisme des défauts observés dans un type de cellule particulier, à un moment particulier du développement.

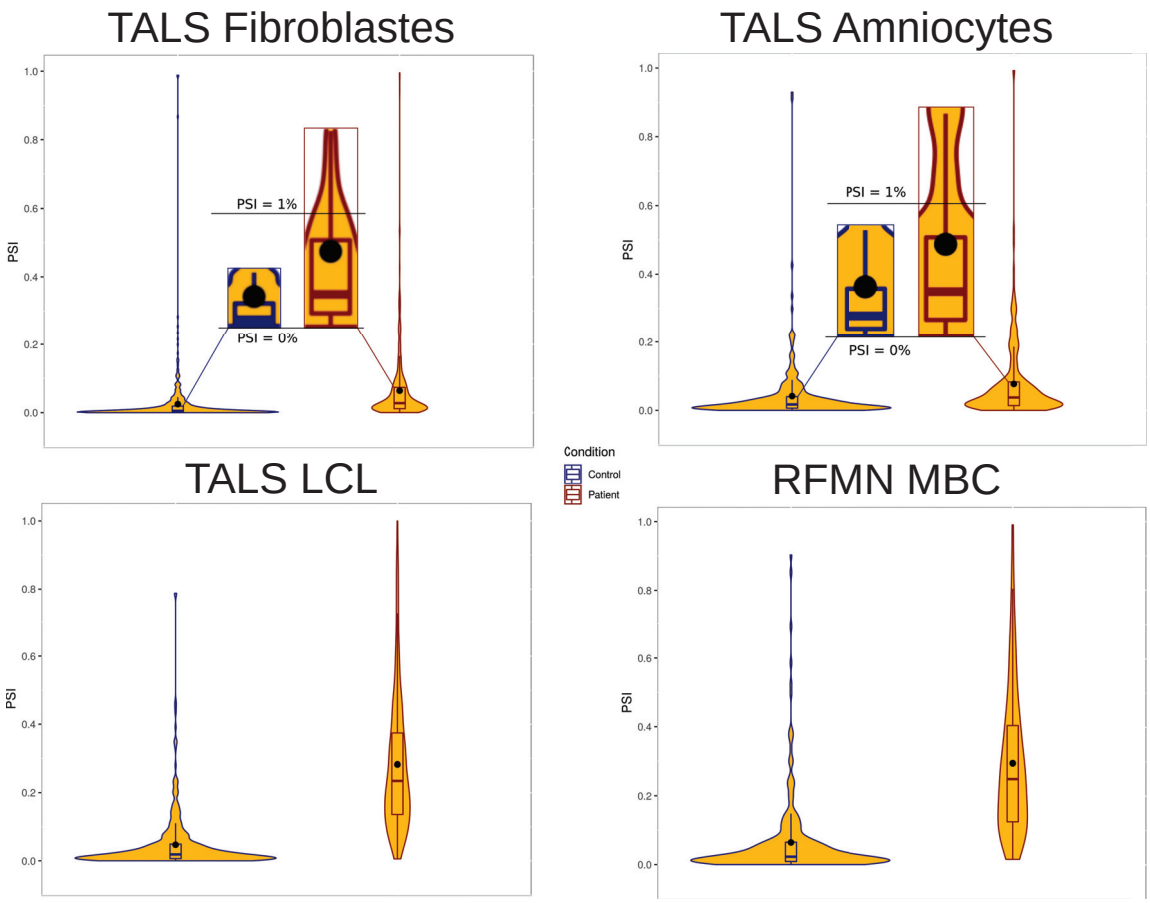

Figure 28 : Distribution des PSI des introns U12 analysés dans les quatre jeux de données.
Boxplots et violinplots des PSI U12 contrôles (bleu) et patients (rouge) TALS fibroblastes, amniocytes, LCL et RFMN MBC. Un point noir indique la moyenne de la distribution. MBC = cellules sanguines mononucléées.

b. Impact biologique des rétentions d'intron U12

Dans le contexte des fibroblastes et des amniocytes de patients TALS, on peut se demander quel peut être l'impact biologique des nombreuses, mais très faibles, rétentions d'introns mineurs observées. Un PSI passant de 20 % chez les contrôles à 21 % chez les patients semble *a priori* avoir moins d'impact qu'un PSI passant de 0 % à 1 %, bien qu'ils représentent tous deux une différence de PSI de 1 %. En effet, dans ce dernier cas, la pression sur la qualité de l'épissage de l'intron parait élevée (intron parfaitement épissé en condition physiologique). Ainsi, une perturbation de l'épissage d'un tel intron pourrait plus probablement avoir des conséquences biologiques. Dans le futur, il serait intéressant de calculer des Fold Change de PSI (PSI_patient / PSI_contrôle) pour avoir une métrique qui prendrait en compte la sensibilité d'un intron à une variation minime de la qualité de son épissage.

Paradoxalement, l'expression des gènes U2 et U12 est similaire entre patients TALS et contrôles, alors que l'on pourrait attendre que les transcrits contenant un intron U12 soient retenus dans le noyau ou dégradés par des mécanismes de contrôle qualité nucléaire et/ou cytoplasmique (Verma et al., 2018), tous deux pouvant mener, en principe, à une sous-expression des gènes U12 si aucun mécanisme de feed-back n'est en action. Les transcrits aberrants semblent vraisemblablement échapper à ces contrôles qualités, par un mécanisme qui reste encore à élucider. Bien qu'il soit impossible, avec nos données de séquençage, de prédire l'impact sur le niveau de protéine fonctionnelle, il semble probable que ce niveau soit fortement réduit, les rétentions d'introns introduisant presque toujours au moins un codon stop prématuré dans le cadre de lecture des transcrits, même si le niveau d'expression du gène est inchangé. Des études sur le niveau d'expression des protéines seraient importantes à mettre en place, pour mieux répondre à la question des conséquences associées aux rétentions d'intron.

c. Utilisation préférentielle de sites d'épissage U2

Mis à part les IR, d'autres événements d'épissage intéressants ont pu être détectés chez les patients, comme des rétentions d'introns U2 voisins aux introns U12 retenus, indiquant vraisemblablement une interaction entre les spliceosomes mineur et majeur, ou l'utilisation de sites cryptiques U2 aux alentours des sites U12, particulièrement fréquents dans les LCL. La quasi-totalité de ces derniers événements conduisent à un changement de site donneur (altD), couplé parfois avec un changement de site accepteur (altAD, illustré par des traits noirs dans la Figure 27). Ce changement d'un site donneur U12 vers un site donneur U2 voisin peut être expliquée par la forte spécificité des séquences de chaque type de site donneur, contrairement aux sites accepteurs mineurs et majeurs qui sont presque identiques (faible spécificité). Les altAD détectés par KisSplice semblent donc abondants, pourtant il s'agit d'un événement particulièrement complexe, dont seulement une partie est détectable par KisSplice, alors que la plupart des outils d'analyse de l'épissage ne les recherchent pas. Il est probable que beaucoup d'événements de donneurs/accepteurs alternatifs soient actuellement indétectables, et un développement méthodologique de KisSplice et/ou KisSplice2RefGenome, que nous envisageons, pourrait améliorer leur détection. Il serait aussi possible de développer une méthode basée sur l'alignement pour répondre à ce problème bioinformatique. J'envisage ainsi de contribuer à la détection des altAD en modifiant certaines fonctionnalités de KisSplice et/ou en développant une méthode dédiée.

d. Reclassification d'intron U2 en U12

En nous intéressant aux rétentions d'introns U2 chez les patients TALS, nous avons pu déterminer que certains de ces introns pouvaient être considérés comme des introns U12 en se basant sur les scores associés à leurs sites d'épissage U12 et U2, ainsi que sur la force de la rétention de l'intron dans les données TALS. Notre jeu de données peut donc être utilisé comme guide pour clarifier le type des introns situés à la frontière entre mineur et majeur.

De plus, il semble probable que l'ensemble des introns U12, connus ou non, situés à l'intérieur des gènes exprimés dans les échantillons LCL de patients TALS soient retenus. Ce jeu de données pourrait ainsi être utilisé pour définir une nouvelle et large liste de « vrais » intron U12, sur laquelle des réseaux de neurones pourraient être entraînés pour ensuite déterminer si d'autres introns U12 sont situés ailleurs dans le génome, et plus particulièrement dans des gènes non-exprimés chez les TALS LCL.

e. Forte couverture en reads de gènes de snRNA

Une autre caractéristique du jeu de données correspondant au patient TALS LCL est la multitude de reads s'alignant sur les gènes de snRNA mineurs et majeurs, à l'exception de U6 et U6atac. Cette observation, non retrouvée dans les fibroblastes et amniocytes de patients TALS, laisse penser que les snRNA sont polyadénylés (puisqu'ils sont séquencés suite à une capture des queues poly-A), suite à un problème au niveau de leur maturation. De nombreux gènes U12 codant pour des protéines compose le complexe Integrator, nécessaire à cette maturation. Nous avons donc émis l'hypothèse que ce complexe Integrator pourrait être impliqué dans ce phénomène. Il semble cependant improbable que ce défaut de maturation soit à lui seul l'origine des fortes rétentions U12 observées dans les LCL, puisque l'épissage des introns U2 est globalement épargné dans ces cellules, bien que les snRNA majeurs soient eux aussi affectés. Ce défaut de maturation pourrait engendrer le stockage des snRNA dans une structure du cytoplasme riche en snRNP, les U-bodies (Liu and Gall, 2007), et ainsi mener au séquençage massif des snRNA autre que U6 et U6atac. Les U-bodies sont des foyers cytoplasmiques riches en U-snRNP, qui pourraient être des points de passage lors de la maturation des snRNA, juste avant leur importation dans le noyau, puisque le complexe SMN est lui aussi localisé dans les U-bodies.

f. TALS et déficit immunitaire

Une autre observation soutenant l'hypothèse que les U-bodies peuvent-être anormaux chez les patients TALS est le fait qu'ils semblent impliqués dans la réponse immunitaire. En effet, la

quantité physiologique de U-bodies est dépendante du niveau de stress de la cellule et est modifiée lors d'une carence en acides aminés causée, par exemple, par une infection bactérienne (Tsalikis et al., 2015). Il se trouve aussi que les patients TALS semblent particulièrement sensibles aux infections, le décès arrivant le plus souvent dans les premières années de la vie post-natale, après un épisode infectieux anodin.

Suite à nos résultats montrant que les gènes U12 les plus fortement impactés chez les TALS sont associés à des processus immunitaires, notre équipe a pu observer une déficience immunitaire chez les TALS, semblable à celle de patients RFMN (résultat non-publié). Il est aussi intéressant d'observer que, chez les fœtus de femmes affectés par le virus Zika, qui présentent eux aussi un nanisme microcéphalique sévère, la réponse immunitaire pourrait provoquer une désorientation du réseau mitotique des cellules, réseau essentiel pour la division asymétrique qui permettra le développement des organes et la différentiation des cellules (McDougall et al., 2019). Tous ces indices indiquent que la réponse immunitaire des cellules pourrait être un élément important de la physiopathologie du TALS.

C. Perspectives

Cependant, nous ne connaissons toujours pas la chaîne de causalité permettant de lier l'épissage mineur et le développement embryonnaire/cérébrale. Le profil transcriptomique est très variable au cours du développement, et l'accumulation de petits défauts d'épissage mineur après la différentiation des cellules neurales pourrait conduire à un dysfonctionnement ou à la mort des neurones (Jutzi et al., 2018).

a. BrainSpan : étude du cerveau au cours du développement

Dans le but d'en apprendre davantage sur le profil d'épissage mineur au cours du développement cérébral, nous avons fait une demande d'accès aux données brutes générées par le projet BrainSpan (Miller et al., 2014), récemment acceptée. Dans ce projet, plus de 600 RNA-seq de diverses parties du cerveau à divers moments du développement du fœtus et de l'enfant ont été réalisés. L'analyse de ces données devrait nous donner un aperçu de la qualité de l'épissage mineur physiologique dans le cerveau durant le développement. Si des introns U12 sont particulièrement bien épissés ou si des gènes U12 ne sont exprimés qu'à certains moments cruciaux du développement cérébral, alors ils pourraient être particulièrement sensibles à une rétention et donc importants pour comprendre la physiopathologie du TALS.

b. Le poisson-zèbre : modèle animal pour le TALS

Pour comprendre le développement des cellules neurales de patients TALS, notre équipe s'est dirigée vers la mise au point d'un modèle animal, le poisson-zèbre, un excellent modèle d'étude de l'organogenèse des vertébrés, pour poursuivre l'étude du rôle joué par l'épissage mineur lors du développement, avec une attention particulière pour le neurodéveloppement. Un knock-down du gène *u6atac* à l'aide de morpholinos (oligomères utilisés pour réduire l'expression d'un gène) dans des embryons de cet organisme ainsi que dans des cellules de mammifère démontrent un rôle conservé du spliceosome mineur dans le cycle cellulaire (König et al., 2007) et l'embryogenèse (Markmiller et al., 2014). Nous avons effectué un knock-out du gène *u4atac* du poisson-zèbre en 2016, grâce à un financement de la Fondation Maladies Rares, et la génération F1 est actuellement en cours d'élevage. En attendant, l'équipe utilise des morpholinos pour vérifier qu'une perte de fonction de *u4atac* chez le poisson-zèbre conduit bien aux défauts cliniques du TALS.

L'équipe envisage également la possibilité de travailler avec des cellules souches pluripotentes induites humaines.

c. TALS et ciliopathie

En parallèle de cette thèse, de nouveaux patients atteints du TALS, mais ne présentant aucune mutation dans le gène *RNU4ATAC*, ont été identifiés. L'un de ces patients présente une mutation homozygote dans le gène *RTTN*, codant pour une protéine associée au cil primaire (antenne cellulaire à l'interface des milieux intra- et extra-cellulaire) et au centrosome (centre de l'organisation des réseaux de microtubules, impliqué dans la division cellulaire). Aucune conséquence sur l'épissage des introns U12 n'a été détectée chez ce patient (Figure 29). Nous n'avons pas pu identifier un lien direct entre *U4atac* et *RTTN*. Bien que ce dernier disposait d'un intron U12/U2 dans le base de données U12db, datant de 2007, rien dans nos résultats ne nous incite à considérer cet intron comme mineur, puisqu'il a été reclassifié en intron U2 dans notre version de l'annotation et qu'il n'est pas retenu dans les cellules de patients TALS ou RFMN.

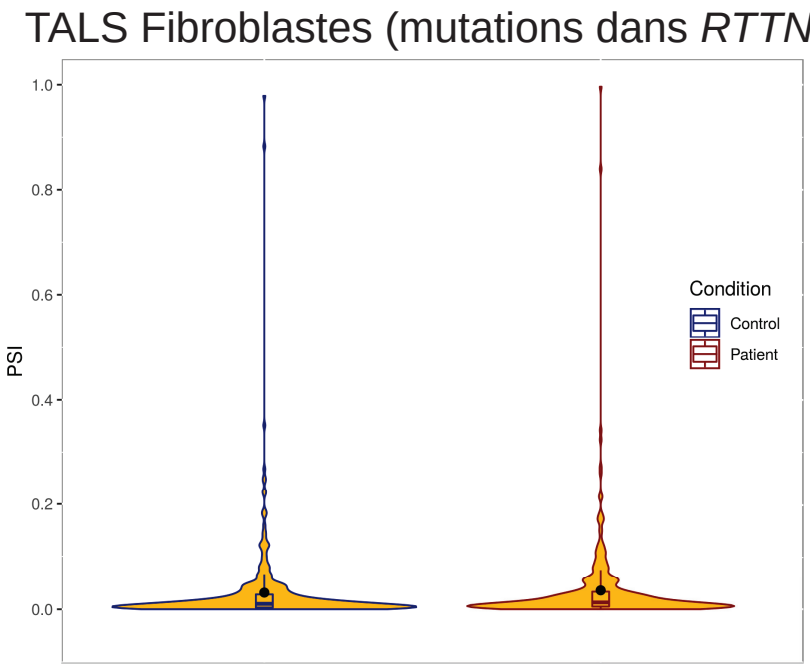

Figure 29 : Qualité de l'épissage des introns U12 chez les patients TALS avec mutations dans *RTTN*.
Boxplots et violinplots des PSI U12 contrôles (bleu) et patients (rouge) TALS fibroblastes sans mutation dans le gène *RNU4ATAC* mais avec mutations dans le gène *RTTN*. Un point noir indique la moyenne de la distribution.

De plus, deux cas de Syndrome de Joubert (JBS), une maladie liée au centrosome/cil primaire (ciliopathie), provoquée par une mutation homozygote de *RNU4ATAC* nous ont été communiqués. En consultant spécifiquement des bases de données recensant les gènes du cil, nous avons aussi découvert que les gènes U12 particulièrement affectés dans les cellules de patients TALS étaient enrichis pour la biogenèse du cil primaire. En 2018, l'ensemble de ces découvertes, non publiées, a permis à notre équipe de recevoir un financement de l'Agence Nationale de la Recherche pour continuer nos études sur le poisson-zèbre, le cil primaire, les patients TALS *RTTN* et les patients JBS *RNU4ATAC*. Les objectifs sont alors de comprendre les relations génotype-phénotype du TALS, déterminer si les maladies liées à *RNU4ATAC* provoquent une atteinte du cil primaire, découvrir comment l'épissage mineur ou le snRNA U4atac pourraient être reliés au cil primaire et analyser les transcriptomes de poisson-zèbres et de nouveaux patients TALS. C'est dans ce projet que s'inscrirait le futur de ma carrière professionnelle, une place de post-doctorant m'étant proposée durant les deux prochaines années, pour mener l'analyse bioinformatique des multiples séquençages d'ARN qui seront réalisés, à la fois chez des individus et des poisson-zèbres mutants en cours de développement.

Sources

Abdel-Salam, G.M.H., Abdel-Hamid, M.S., Issa, M., Magdy, A., El-Kotoury, A., Amr, K., 2012. Expanding the phenotypic and mutational spectrum in microcephalic osteodysplastic primordial dwarfism type I. American Journal of Medical Genetics Part A 158A, 1455–1461. https://doi.org/10.1002/ajmg.a.35356

Alekseyenko, A.V., Kim, N., Lee, C.J., 2007. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA 13, 661–670. https://doi.org/10.1261/rna.325107

Alioto, T.S., 2007. U12DB: a database of orthologous U12-type spliceosomal introns. Nucleic Acids Research 35, D110–D115. https://doi.org/10.1093/nar/gkl796

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169. https://doi.org/10.1093/bioinformatics/btu638

Anna, A., Monika, G., 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. Journal of Applied Genetics 59, 253–268. https://doi.org/10.1007/s13353-018-0444-7

Argente, J., Flores, R., Gutierrez-Arumi, A., Verma, B., Martos-Moreno, G.A., Cusco, I., Oghabian, A., Chowen, J.A., Frilander, M.J., Perez-Jurado, L.A., 2014. Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. EMBO Molecular Medicine 6, 299–306. https://doi.org/10.1002/emmm.201303573

Baillat, D., Hakimi, M.-A., Näär, A.M., Shilatifard, A., Cooch, N., Shiekhattar, R., 2005. Integrator, a Multiprotein Mediator of Small Nuclear RNA Processing, Associates with the C-Terminal Repeat of RNA Polymerase II. Cell 123, 265–276. https://doi.org/10.1016/j.cell.2005.08.019

Barash, Y., Blencowe, B.J., Frey, B.J., 2010. Model-based detection of alternative splicing signals. Bioinformatics 26, i325–i333. https://doi.org/10.1093/bioinformatics/btq200

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C.M., Wilson, M.D., Kim, P.M., Odom, D.T., Frey, B.J., Blencowe, B.J., 2012. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science 338, 1587–1593. https://doi.org/10.1126/science.1230612

Bartschat, S., Samuelsson, T., 2010. U12 type introns were lost at multiple occasions during evolution. BMC Genomics 11, 106. https://doi.org/10.1186/1471-2164-11-106

Basu, M.K., Makalowski, W., Rogozin, I.B., Koonin, E.V., 2008. U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. Biology Direct 3, 19. https://doi.org/10.1186/1745-6150-3-19

Benoit-Pilven, C., Marchet, C., Chautard, E., Lima, L., Lambert, M.-P., Sacomoto, G., Rey, A., Cologne, A., Terrone, S., Dulaurier, L., Claude, J.-B., Bourgeois, C.F., Auboeuf, D., Lacroix, V., 2018. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. Scientific Reports 8, 4307. https://doi.org/10.1038/s41598-018-21770-7

Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA 74, 3171–3175. https://doi.org/10.1073/pnas.74.8.3171

Bogaert, D.J., Dullaers, M., Kuehn, H.S., Leroy, B.P., Niemela, J.E., De Wilde, H., De Schryver, S., De Bruyne, M., Coppieters, F., Lambrecht, B.N., De Baets, F.,

Rosenzweig, S.D., De Baere, E., Haerynck, F., 2017. Early-onset primary antibody deficiency resembling common variable immunodeficiency challenges the diagnosis of Wiedeman-Steiner and Roifman syndromes. Scientific Reports 7, 3702. https://doi.org/10.1038/s41598-017-02434-4

Boutz, P.L., Bhutkar, A., Sharp, P.A., 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes & Development 29, 63–80. https://doi.org/10.1101/gad.247361.114

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Cossa, V.T.-, Wanunu, M., Wiggin, M., Schloss, J.A., 2009. The potential and challenges of nanopore sequencing, in: Nanoscience and Technology. Co-Published with Macmillan Publishers Ltd, UK, pp. 261–268. https://doi.org/10.1142/9789814287005_0027

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., Blencowe, B.J., 2014. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Research 24, 1774–1786. https://doi.org/10.1101/gr.177790.114

Brock, J.E., Dietrich, R.C., Padgett, R.A., 2008. Mutational analysis of the U12-dependent branch site consensus sequence. RNA 14, 2430–2439. https://doi.org/10.1261/rna.1189008

Burge, C.B., Padgett, R.A., Sharp, P.A., 1998. Evolutionary Fates and Origins of U12-Type Introns. Molecular Cell 2, 773–785. https://doi.org/10.1016/S1097-2765(00)80292-0

Chang, S.Y., Yong, T.F., Yu, C.Y., Liang, M.C., Pletnikova, O., Troncoso, J., Burgunder, J.-M., Soong, T.W., 2007. Age and gender-dependent alternative splicing of P/Q-type calcium channel EF-hand. Neuroscience 145, 1026–1036. https://doi.org/10.1016/j.neuroscience.2006.12.054

Clara Benoit-Pilven, C.M., 2018. kissDE. Bioconductor. https://doi.org/10.18129/b9.bioc.kissde

Cvitkovic, I., Jurica, M.S., 2013. Spliceosome Database: a tool for tracking components of the spliceosome. Nucleic Acids Research 41, D132–D141. https://doi.org/10.1093/nar/gks999

Di, Y., Schafer, D.W., Cumbie, J.S., Chang, J.H., 2011. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology 10. https://doi.org/10.2202/1544-6115.1637

Dietrich, R.C., 2005. A mutational analysis of U12-dependent splice site dinucleotides. RNA 11, 1430–1440. https://doi.org/10.1261/rna.7206305

Dietrich, R.C., Incorvaia, R., Padgett, R.A., 1997. Terminal Intron Dinucleotide Sequences Do Not Distinguish between U2- and U12-Dependent Introns. Molecular Cell 1, 151–160. https://doi.org/10.1016/S1097-2765(00)80016-7

Dietrich, R.C., Peris, M.J., Seyboldt, A.S., Padgett, R.A., 2001. Role of the 3' Splice Site in U12-Dependent Intron Splicing. Molecular and Cellular Biology 21, 1942–1952. https://doi.org/10.1128/MCB.21.6.1942-1952.2001

Dinur Schejter, Y., Ovadia, A., Alexandrova, R., Thiruvahindrapuram, B., Pereira, S.L., Manson, D.E., Vincent, A., Merico, D., Roifman, C.M., 2017. A homozygous mutation in the stem II domain of RNU4ATAC causes typical Roifman syndrome. npj Genomic Medicine 2, 23. https://doi.org/10.1038/s41525-017-0024-5

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Edery, P., Marcaillou, C., Sahbatou, M., Labalme, A., Chastang, J., Touraine, R., Tubacher, E., Senni, F., Bober, M.B., Nampoothiri, S., Jouk, P.-S., Steichen, E., Berland, S., Toutain, A., Wise, C.A., Sanlaville, D., Rousseau, F., Clerget-Darpoux, F., Leutenegger, A.-L., 2011. Association of TALS Developmental Disorder with Defect in Minor Splicing Component U4atac snRNA. Science 332, 240–243. https://doi.org/10.1126/science.1202205

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133–138. https://doi.org/10.1126/science.1162986

Elsaid, M.F., Chalhoub, N., Ben-Omran, T., Kumar, P., Kamel, H., Ibrahim, K., Mohamoud, Y., Al-Dous, E., Al-Azwani, I., Malek, J.A., Suhre, K., Ross, M.E., Aleem, A.A., 2017. Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia: *RNU12* in Cerebellar Ataxia. Annals of Neurology 81, 68–78. https://doi.org/10.1002/ana.24826

Farach, L.S., Little, M.E., Duker, A.L., Logan, C.V., Jackson, A., Hecht, J.T., Bober, M., 2018. The expanding phenotype of *RNU4ATAC* pathogenic variants to Lowry Wood syndrome. American Journal of Medical Genetics Part A 176, 465–469. https://doi.org/10.1002/ajmg.a.38581

Ferrell, S., Johnson, A., Pearson, W., 2016. Microcephalic osteodysplastic primordial dwarfism type 1. BMJ Case Reports bcr2016215502. https://doi.org/10.1136/bcr-2016-215502

Friend, K., Kolev, N.G., Shu, M.-D., Steitz, J.A., 2008. Minor-class splicing occurs in the nucleus of the Xenopus oocyte. RNA 14, 1459–1462. https://doi.org/10.1261/rna.1119708

Frilander, M.J., 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.

Gao, K., Masuda, A., Matsuura, T., Ohno, K., 2008. Human branch point consensus sequence is yUnAy. Nucleic Acids Research 36, 2257–2267. https://doi.org/10.1093/nar/gkn073

Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods 8, 469–477. https://doi.org/10.1038/nmeth.1613

Ge, Y., Porse, B.T., 2014. The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression: Prospects & Overviews. BioEssays 36, 236–243. https://doi.org/10.1002/bies.201300156

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652. https://doi.org/10.1038/nbt.1883

Gruss, O.J., Meduri, R., Schilling, M., Fischer, U., 2017. UsnRNP biogenesis: mechanisms and regulation. Chromosoma 126, 577–593. https://doi.org/10.1007/s00412-017-0637-6

Hallermayr, A., Graf, J., Koehler, U., Laner, A., Schönfeld, B., Benet-Pagès, A., Holinski-Feder, E., 2018. Extending the critical regions for mutations in the non-coding gene

RNU4ATAC in another patient with Roifman Syndrome. Clinical Case Reports 6, 2224–2228. https://doi.org/10.1002/ccr3.1830

Hansen, K.D., Brenner, S.E., Dudoit, S., 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 38, e131–e131. https://doi.org/10.1093/nar/gkq224

Hardiman, O., Al-Chalabi, A., Brayne, C., Beghi, E., van den Berg, L.H., Chio, A., Martin, S., Logroscino, G., Rooney, J., 2017. The changing picture of amyotrophic lateral sclerosis: lessons from European registers. Journal of Neurology, Neurosurgery & Psychiatry 88, 557–563. https://doi.org/10.1136/jnnp-2016-314495

He, H., Liyanarachchi, S., Akagi, K., Nagy, R., Li, J., Dietrich, R.C., Li, W., Sebastian, N., Wen, B., Xin, B., Singh, J., Yan, P., Alder, H., Haan, E., Wieczorek, D., Albrecht, B., Puffenberger, E., Wang, H., Westman, J.A., Padgett, R.A., Symer, D.E., de la Chapelle, A., 2011. Mutations in U4atac snRNA, a Component of the Minor Spliceosome, in the Developmental Disorder MOPD I. Science 332, 238–240. https://doi.org/10.1126/science.1200587

Heremans, J., Garcia-Perez, J.E., Turro, E., Schlenner, S.M., Casteels, I., Collin, R., de Zegher, F., Greene, D., Humblet-Baron, S., Lesage, S., Matthys, P., Penkett, C.J., Put, K., Stirrups, K., Thys, C., Van Geet, C., Van Nieuwenhove, E., Wouters, C., Meyts, I., Freson, K., Liston, A., 2018. Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. Journal of Allergy and Clinical Immunology 142, 630–646. https://doi.org/10.1016/j.jaci.2017.11.061

Heyn, P., Kalinka, A.T., Tomancak, P., Neugebauer, K.M., 2015. Introns and gene expression: Cellular constraints, transcriptional regulation, and evolutionary consequences: Prospects & Overviews. BioEssays 37, 148–154. https://doi.org/10.1002/bies.201400138

Hintze, J.L., Nelson, R.D., 1998. Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician 52, 181. https://doi.org/10.2307/2685478

Horiuchi, K., Perez-Cerezales, S., Papasaikas, P., Ramos-Ibeas, P., López-Cardona, A.P., Laguna-Barraza, R., Fonseca Balvís, N., Pericuesta, E., Fernández-González, R., Planells, B., Viera, A., Suja, J.A., Ross, P.J., Alén, F., Orio, L., Rodriguez de Fonseca, F., Pintado, B., Valcárcel, J., Gutiérrez-Adán, A., 2018. Impaired Spermatogenesis, Muscle, and Erythrocyte Function in U12 Intron Splicing-Defective Zrsr1 Mutant Mice. Cell Reports 23, 143–155. https://doi.org/10.1016/j.celrep.2018.03.028

Irimia, M., Roy, S.W., 2014. Origin of Spliceosomal Introns and Alternative Splicing. Cold Spring Harbor Perspectives in Biology 6, a016071–a016071. https://doi.org/10.1101/cshperspect.a016071

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., Barrios-Rodiles, M., Sternberg, M.J.E., Cordes, S.P., Roth, F.P., Wrana, J.L., Geschwind, D.H., Blencowe, B.J., 2014. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell 159, 1511–1523. https://doi.org/10.1016/j.cell.2014.11.035

Jackson, IJ, 1991. A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Research 19, 3795–3798.

Jacob, A.G., Smith, C.W.J., 2017. Intron retention as a component of regulated gene expression programs. Human Genetics 136, 1043–1057. https://doi.org/10.1007/s00439-017-1791-x

Jafarifar, F., Dietrich, R.C., Hiznay, J.M., Padgett, R.A., 2014. Biochemical defects in minor spliceosome function in the developmental disorder MOPD I. RNA 20, 1078–1089. https://doi.org/10.1261/rna.045187.114

Jia, Y., Mu, J.C., Ackerman, S.L., 2012. Mutation of a U2 snRNA Gene Causes Global Disruption of Alternative Splicing and Neurodegeneration. Cell 148, 296–308. https://doi.org/10.1016/j.cell.2011.11.057

Jutzi, D., Akinyi, M.V., Mechtersheimer, J., Frilander, M.J., Ruepp, M.-D., 2018. The emerging role of minor intron splicing in neurological disorders. Cell Stress 2, 40–54. https://doi.org/10.15698/cst2018.03.126

Katz, Y., Wang, E.T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., Airoldi, E.M., Burge, C.B., 2015. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics 31, 2400–2402. https://doi.org/10.1093/bioinformatics/btv034

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.-H.A., Marr, M.T., Rosbash, M., 2011. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes & Development 25, 2502–2512. https://doi.org/10.1101/gad.178962.111

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods 12, 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14, R36. https://doi.org/10.1186/gb-2013-14-4-r36

König, H., Matter, N., Bader, R., Thiele, W., Müller, F., 2007. Splicing Segregation: The Minor Spliceosome Acts outside the Nucleus and Controls Cell Proliferation. Cell 131, 718–729. https://doi.org/10.1016/j.cell.2007.09.043

Konig, H., Muller, F., 2008. Minor splicing: Nuclear dogma still in question. Proceedings of the National Academy of Sciences 105, E37–E37. https://doi.org/10.1073/pnas.0804939105

Kumar, A., 2009. An Overview of Nested Genes in Eukaryotic Genomes. Eukaryotic Cell 8, 1321–1329. https://doi.org/10.1128/EC.00143-09

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lardelli, R.M., Schaffer, A.E., Eggens, V.R.C., Zaki, M.S., Grainger, S., Sathe, S., Van Nostrand, E.L., Schlachetzki, Z., Rosti, B., Akizu, N., Scott, E., Silhavy, J.L., Heckman, L.D., Rosti, R.O., Dikoglu, E., Gregor, A., Guemez-Gamboa, A., Musaev, D., Mande, R., Widjaja, A., Shaw, T.L., Markmiller, S., Marin-Valencia, I., Davies, J.H., de Meirleir, L., Kayserili, H., Altunoglu, U., Freckmann, M.L., Warwick, L., Chitayat, D., Blaser, S., Çağlayan, A.O., Bilguvar, K., Per, H., Fagerberg, C., Christesen, H.T., Kibaek, M., Aldinger, K.A., Manchester, D., Matsumoto, N., Muramatsu, K., Saitsu, H., Shiina, M., Ogata, K., Foulds, N., Dobyns, W.B., Chi, N.C., Traver, D., Spaccini, L., Bova, S.M., Gabriel, S.B., Gunel, M., Valente, E.M., Nassogne, M.-C., Bennett, E.J., Yeo, G.W., Baas, F., Lykke-Andersen, J., Gleeson, J.G., 2017. Biallelic mutations in the 3′ exonuclease TOE1 cause pontocerebellar hypoplasia and uncover a role in snRNA processing. Nature Genetics 49, 457–464. https://doi.org/10.1038/ng.3762

Lareau, L.F., Brooks, A.N., Soergel, D.A.W., Meng, Q., Brenner, S.E., 2007. The Coupling of Alternative Splicing and Nonsense-Mediated mRNA Decay, in: Blencowe, B.J., Graveley, B.R. (Eds.), Alternative Splicing in the Postgenomic Era. Springer New York, New York, NY, pp. 190–211. https://doi.org/10.1007/978-0-387-77374-2_12

Le Hir, H., 2001. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. The EMBO Journal 20, 4987–4997. https://doi.org/10.1093/emboj/20.17.4987

Le Hir, H., Nott, A., Moore, M.J., 2003. How introns influence and enhance eukaryotic gene expression. Trends in Biochemical Sciences 28, 215–220. https://doi.org/10.1016/S0968-0004(03)00052-5

Levine, A., Durbin, R., 2001. A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Research 29, 4006–4013. https://doi.org/10.1093/nar/29.19.4006

Lev-Maor, G., 2003. The Birth of an Alternatively Spliced Exon: 3' Splice-Site Selection in Alu Exons. Science 300, 1288–1291. https://doi.org/10.1126/science.1082588

Lewandowska, D., Simpson, C.G., Clark, G.P., Jennings, N.S., Barciszewska-Pacak, M., Lin, C.-F., Makalowski, W., Brown, J.W.S., Jarmolowski, A., 2004. Determinants of Plant U12-Dependent Intron Splicing Efficiency. Plant Cell 16, 1340–1352. https://doi.org/10.1105/tpc.020743

Lewis, B.P., Green, R.E., Brenner, S.E., 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proceedings of the National Academy of Sciences 100, 189–192. https://doi.org/10.1073/pnas.0136770100

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., Pritchard, J.K., 2016. LeafCutter: Annotation-free quantification of RNA splicing (preprint). Genomics. https://doi.org/10.1101/044107

Lin, C.-F., Mount, S.M., Jarmołowski, A., Makałowski, W., 2010. Evolutionary dynamics of U12-type spliceosomal introns. BMC Evolutionary Biology 10, 47. https://doi.org/10.1186/1471-2148-10-47

Lin, M., Lucas, H.C., Shmueli, G., 2013. **Research Commentary** —Too Big to Fail: Large Samples and the *p* -Value Problem. Information Systems Research 24, 906–917. https://doi.org/10.1287/isre.2013.0480

Lionel, A.C., Costain, G., Monfared, N., Walker, S., Reuter, M.S., Hosseini, S.M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W.W.L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M.T., Cordeiro, D., Cytrynbaum, C., Dell, S.D., Dhir, P., Dowling, J.J., Heon, E., Hewson, S., Hiraki, L., Inbar-Feigenberg, M., Klatt, R., Kronick, J., Laxer, R.M., Licht, C., MacDonald, H., Mercimek-Andrews, S., Mendoza-Londono, R., Piscione, T., Schneider, R., Schulze, A., Silverman, E., Siriwardena, K., Snead, O.C., Sondheimer, N., Sutherland, J., Vincent, A., Wasserman, J.D., Weksberg, R., Shuman, C., Carew, C., Szego, M.J., Hayeems, R.Z., Basran, R., Stavropoulos, D.J., Ray, P.N., Bowdin, S., Meyn, M.S., Cohn, R.D., Scherer, S.W., Marshall, C.R., 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genetics in Medicine 20, 435–443. https://doi.org/10.1038/gim.2017.119

Liu, J.-L., Gall, J.G., 2007. U bodies are cytoplasmic structures that contain uridine-rich small nuclear ribonucleoproteins and associate with P bodies. Proceedings of the National Academy of Sciences 104, 11655–11659. https://doi.org/10.1073/pnas.0704977104

Liu, Y., Ferguson, J.F., Xue, C., Silverman, I.M., Gregory, B., Reilly, M.P., Li, M., 2013. Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose. PLoS ONE 8, e66883. https://doi.org/10.1371/journal.pone.0066883

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, Z., Matera, A.G., 2015. Developmental Analysis of Spliceosomal snRNA Isoform Expression. G3: Genes|Genomes|Genetics 5, 103–110. https://doi.org/10.1534/g3.114.015735

Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., Sanada, M., Grossmann, V., Sundaresan, J., Shiraishi, Y., Miyano, S., Thol, F., Ganser, A., Yang, H., Haferlach, T., Ogawa, S., Koeffler, H.P., 2015. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. Nature Communications 6, 6042. https://doi.org/10.1038/ncomms7042

Markmiller, S., Cloonan, N., Lardelli, R.M., Doggett, K., Keightley, M.-C., Boglev, Y., Trotter, A.J., Ng, A.Y., Wilkins, S.J., Verkade, H., Ober, E.A., Field, H.A., Grimmond, S.M., Lieschke, G.J., Stainier, D.Y.R., Heath, J.K., 2014. Minor class splicing shapes the zebrafish transcriptome during development. Proceedings of the National Academy of Sciences 111, 3062–3067. https://doi.org/10.1073/pnas.1305536111

Mauger, O., Lemoine, F., Scheiffele, P., 2016. Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. Neuron 92, 1266–1278. https://doi.org/10.1016/j.neuron.2016.11.032

McDonald, J.H., 2015. Multiple comparisons. Sparky House Publishing Handbook of Biological Statistics (3rd ed.), 254–260.

McDougall, W.M., Perreira, J.M., Hung, H.-F., Vertii, A., Xiaofei, E., Zimmerman, W., Kowalik, T.F., Doxsey, S., Brass, A.L., 2019. Viral Infection or IFN-α Alters Mitotic Spindle Orientation by Modulating Pericentrin Levels. iScience 12, 270–279. https://doi.org/10.1016/j.isci.2019.01.025

McGill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of Box Plots. The American Statistician 32, 12. https://doi.org/10.2307/2683468

Merico, D., Roifman, M., Braunschweig, U., Yuen, R.K.C., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B., Gray, P., Kakakios, A., Peake, J., Hogarth, S., Manson, D., Buncic, R., Pereira, S.L., Herbrick, J.-A., Blencowe, B.J., Roifman, C.M., Scherer, S.W., 2015. Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. Nature Communications 6, 8718. https://doi.org/10.1038/ncomms9718

Middleton, R., Gao, D., Thomas, A., Singh, B., Au, A., Wong, J.J.-L., Bomane, A., Cosson, B., Eyras, E., Rasko, J.E.J., Ritchie, W., 2017. IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biology 18, 51. https://doi.org/10.1186/s13059-017-1184-4

Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., Arnold, J.M., Bennet, C., Bertagnolli, D., Brouner, K., Butler, S., Caldejon, S., Carey, A., Cuhaciyan, C., Dalley, R.A., Dee, N., Dolbeare, T.A., Facer, B.A.C., Feng, D., Fliss, T.P., Gee, G., Goldy, J., Gourley, L., Gregor, B.W., Gu, G., Howard, R.E., Jochim, J.M., Kuan, C.L., Lau, C., Lee, C.-K., Lee, F., Lemon, T.A., Lesnar, P., McMurray, B., Mastan, N., Mosqueda, N., Naluai-Cecchini, T., Ngo, N.-K., Nyhus, J., Oldre, A., Olson, E., Parente, J., Parker, P.D., Parry, S.E., Stevens, A., Pletikos, M., Reding, M., Roll, K., Sandman, D., Sarreal, M., Shapouri, S., Shapovalova, N.V., Shen, E.H., Sjoquist, N., Slaughterbeck, C.R., Smith, M., Sodt, A.J., Williams, D., Zöllei, L., Fischl, B., Gerstein, M.B., Geschwind, D.H., Glass, I.A., Hawrylycz, M.J., Hevner, R.F., Huang, H., Jones, A.R., Knowles, J.A., Levitt, P., Phillips, J.W., Šestan, N., Wohnoutka, P., Dang, C., Bernard, A., Hohmann, J.G., Lein, E.S., 2014. Transcriptional landscape of the prenatal human brain. Nature 508, 199–206. https://doi.org/10.1038/nature13185

Mlinarić, A., Horvat, M., Šupak Smolčić, V., 2017. Dealing with the positive publication bias: Why you should really publish your negative results. Biochemia Medica 27, 030201. https://doi.org/10.11613/BM.2017.030201

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621–628. https://doi.org/10.1038/nmeth.1226

Neuenkirchen, N., Chari, A., Fischer, U., 2008. Deciphering the assembly pathway of Sm-class U snRNPs. FEBS Letters 582, 1997–2003. https://doi.org/10.1016/j.febslet.2008.03.009

Niemela, E.H., Oghabian, A., Staals, R.H.J., Greco, D., Pruijn, G.J.M., Frilander, M.J., 2014. Global analysis of the nuclear processing of transcripts with unspliced U12-type introns by the exosome. Nucleic Acids Research 42, 7358–7369. https://doi.org/10.1093/nar/gku391

Nilsen, T.W., Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463. https://doi.org/10.1038/nature08909

Norppa, A.J., Kauppala, T.M., Heikkinen, H.A., Verma, B., Iwaï, H., Frilander, M.J., 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. https://doi.org/10.1261/rna.062844.117

Nott, A., 2003. A quantitative analysis of intron effects on mammalian gene expression. RNA 9, 607–617. https://doi.org/10.1261/rna.5250403

Oegema, R., Baillat, D., Schot, R., van Unen, L.M., Brooks, A., Kia, S.K., Hoogeboom, A.J.M., Xia, Z., Li, W., Cesaroni, M., Lequin, M.H., van Slegtenhorst, M., Dobyns, W.B., de Coo, I.F.M., Verheijen, F.W., Kremer, A., van der Spek, P.J., Heijsman, D., Wagner, E.J., Fornerod, M., Mancini, G.M.S., 2017. Human mutations in integrator complex subunits link transcriptome integrity to brain development. PLOS Genetics 13, e1006809. https://doi.org/10.1371/journal.pgen.1006809

O'Reilly, D., Dienstbier, M., Cowley, S.A., Vazquez, P., Drozdz, M., Taylor, S., James, W.S., Murphy, S., 2013. Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. Genome Research 23, 281–291. https://doi.org/10.1101/gr.142968.112

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 40, 1413–1415. https://doi.org/10.1038/ng.259

Patel, A.A., 2002. The splicing of U12-type introns can be a rate-limiting step in gene expression. The EMBO Journal 21, 3804–3815. https://doi.org/10.1093/emboj/cdf297

Pessa, H.K.J., 2006. The abundance of the spliceosomal snRNPs is not limiting the splicing of U12-type introns. RNA 12, 1883–1892. https://doi.org/10.1261/rna.213906

Pessa, H.K.J., Will, C.L., Meng, X., Schneider, C., Watkins, N.J., Perala, N., Nymark, M., Turunen, J.J., Luhrmann, R., Frilander, M.J., 2008. Minor spliceosome components are predominantly localized in the nucleus. Proceedings of the National Academy of Sciences 105, 8655–8660. https://doi.org/10.1073/pnas.0803646105

Pickrell, J.K., Pai, A.A., Gilad, Y., Pritchard, J.K., 2010. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. PLoS Genetics 6, e1001236. https://doi.org/10.1371/journal.pgen.1001236

Putoux, A., Alqahtani, A., Pinson, L., Paulussen, A.D.C., Michel, J., Besson, A., Mazoyer, S., Borg, I., Nampoothiri, S., Vasiljevic, A., Uwineza, A., Boggio, D., Champion, F., de Die-Smulders, C.E., Gardeitchik, T., van Putten, W.K., Perez, M.J., Musizzano, Y., Razavi, F., Drunat, S., Verloes, A., Hennekam, R., Guibaud, L., Alix, E., Sanlaville, D., Lesca, G., Edery, P., 2016. Refining the phenotypical and mutational spectrum of Taybi-Linder syndrome: Phenotypical and mutational spectrum of TALS. Clinical Genetics 90, 550–555. https://doi.org/10.1111/cge.12781

Raj, B., Blencowe, B.J., 2015. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron 87, 14–27. https://doi.org/10.1016/j.neuron.2015.05.004

Rearick, D., Prakash, A., McSweeny, A., Shepard, S.S., Fedorova, L., Fedorov, A., 2011. Critical association of ncRNA with introns. Nucleic Acids Research 39, 2357–2366. https://doi.org/10.1093/nar/gkq1080

Robert, C., Watson, M., 2015. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol 16, 177. https://doi.org/10.1186/s13059-015-0734-x

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nature Biotechnology 29, 24–26. https://doi.org/10.1038/nbt.1754

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. Genome Biology 14, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Russell, A.G., Charette, J.M., Spencer, D.F., Gray, M.W., 2006. An early evolutionary origin for the minor spliceosome. Nature 443, 863–866. https://doi.org/10.1038/nature05228

Sacomoto, G.A., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., Lacroix, V., 2012. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics 13 Suppl 6. https://doi.org/10.1186/1471-2105-13-S6-S5

Scamborova, P., Wong, A., Steitz, J.A., 2004. An Intronic Enhancer Regulates Splicing of the Twintron of Drosophila melanogaster prospero Pre-mRNA by Two Different Spliceosomes. Molecular and Cellular Biology 24, 1855–1869. https://doi.org/10.1128/MCB.24.5.1855-1869.2004

Schneider, C., Will, C.L., Makarova, O.V., Makarov, E.M., Luhrmann, R., 2002. Human U4/U6.U5 and U4atac/U6atac.U5 Tri-snRNPs Exhibit Similar Protein Compositions. Molecular and Cellular Biology 22, 3219–3229. https://doi.org/10.1128/MCB.22.10.3219-3229.2002

Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., Barton, G.J., 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22, 839–851. https://doi.org/10.1261/rna.053959.115

Scotti, M.M., Swanson, M.S., 2016. RNA mis-splicing in disease. Nature Reviews Genetics 17, 19–32. https://doi.org/10.1038/nrg.2015.3

Sessegolo, C., Cruaud, C., Da Silva, C., Dubarry, M., Derrien, T., Lacroix, V., Aury, J.-M., 2019. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules (preprint). Bioinformatics. https://doi.org/10.1101/575142

Shabalina, S.A., Ogurtsov, A.Y., Spiridonov, A.N., Novichkov, P.S., Spiridonov, N.A., Koonin, E.V., 2010. Distinct Patterns of Expression and Evolution of Intronless and Intron-Containing Mammalian Genes. Molecular Biology and Evolution 27, 1745–1749. https://doi.org/10.1093/molbev/msq086

Shaheen, R., Maddirevula, S., Ewida, N., Alsahli, S., Abdel-Salam, G.M.H., Zaki, M.S., Tala, S.A., Alhashem, A., Softah, A., Al-Owain, M., Alazami, A.M., Abadel, B., Patel, N., Al-Sheddi, T., Alomar, R., Alobeid, E., Ibrahim, N., Hashem, M., Abdulwahab, F., Hamad, M., Tabarki, B., Alwadei, A.H., Alhazzani, F., Bashiri, F.A., Kentab, A., Şahintürk, S., Sherr, E., Fregeau, B., Sogati, S., Alshahwan, S.A.M., Alkhalifi, S.,

Alhumaidi, Z., Temtamy, S., Aglan, M., Otaify, G., Girisha, K.M., Tulbah, M., Seidahmed, M.Z., Salih, M.A., Abouelhoda, M., Momin, A.A., Saffar, M.A., Partlow, J.N., Arold, S.T., Faqeih, E., Walsh, C., Alkuraya, F.S., 2019. Genomic and phenotypic delineation of congenital microcephaly. Genetics in Medicine 21, 545–552. https://doi.org/10.1038/s41436-018-0140-3

Shelihan, I., Ehresmann, S., Magnani, C., Forzano, F., Baldo, C., Brunetti-Pierri, N., Campeau, P.M., 2018. Lowry-Wood syndrome: further evidence of association with RNU4ATAC, and correlation between genotype and phenotype. Human Genetics 137, 905–909. https://doi.org/10.1007/s00439-018-1950-8

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., Sachidanandam, R., 2006. Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Research 34, 3955–3967. https://doi.org/10.1093/nar/gkl556

Singh, J., Padgett, R.A., 2009. Rates of in situ transcription and splicing in large human genes. Nature Structural & Molecular Biology 16, 1128–1133. https://doi.org/10.1038/nsmb.1666

Soneson, C., Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14, 91. https://doi.org/10.1186/1471-2105-14-91

Sugarman, E.A., Nagan, N., Zhu, H., Akmaev, V.R., Zhou, Z., Rohlfs, E.M., Flynn, K., Hendrickson, B.C., Scholl, T., Sirko-Osadsa, D.A., Allitto, B.A., 2012. Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: clinical laboratory analysis of >72 400 specimens. European Journal of Human Genetics 20, 27–32. https://doi.org/10.1038/ejhg.2011.134

Tapial, J., Ha, K.C.H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Vallières, M., Permanyer, J., Sodaei, R., Marquez, Y., Cozzuto, L., Wang, X., Gómez-Velázquez, M., Rayon, T., Manzanares, M., Ponomarenko, J., Blencowe, B.J., Irimia, M., 2017. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Research 27, 1759–1768. https://doi.org/10.1101/gr.220962.117

Tarn, W.-Y., 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. https://doi.org/10.1016/S0092-8674(00)81057-0

Tenny, S., Abdelgawad, I., 2019. Statistical Significance, in: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

The RGASP Consortium, Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigó, R., Bertone, P., 2013a. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10, 1185–1191. https://doi.org/10.1038/nmeth.2722

The RGASP Consortium, Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P., 2013b. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 10, 1177–1184. https://doi.org/10.1038/nmeth.2714

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578. https://doi.org/10.1038/nprot.2012.016

Tsalikis, J., Tattoli, I., Ling, A., Sorbara, M.T., Croitoru, D.O., Philpott, D.J., Girardin, S.E., 2015. Intracellular Bacterial Pathogens Trigger the Formation of U Small Nuclear RNA Bodies (U Bodies) through Metabolic Stress Induction. J. Biol. Chem. 290, 20904–20918. https://doi.org/10.1074/jbc.M115.659466

Turunen, J.J., Niemelä, E.H., Verma, B., Frilander, M.J., 2013. The significant other: splicing by the minor spliceosome: Splicing by the minor spliceosome. Wiley Interdisciplinary Reviews: RNA 4, 61–76. https://doi.org/10.1002/wrna.1141

Vanichkina, D.P., Schmitz, U., Wong, J.J.-L., Rasko, J.E.J., 2018. Challenges in defining the role of intron retention in normal biology and disease. Seminars in Cell & Developmental Biology 75, 40–49. https://doi.org/10.1016/j.semcdb.2017.07.030

Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., Barash, Y., 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife 5, e11752. https://doi.org/10.7554/eLife.11752

Vazquez-Arango, P., O'Reilly, D., 2018. Variant snRNPs: New players within the spliceosome system. RNA Biology 15, 17–25. https://doi.org/10.1080/15476286.2017.1373238

Verbeeren, J., Verma, B., Niemelä, E.H., Yap, K., Makeyev, E.V., Frilander, M.J., 2017. Alternative exon definition events control the choice between nuclear retention and cytoplasmic export of U11/U12-65K mRNA. PLOS Genetics 13, e1006824. https://doi.org/10.1371/journal.pgen.1006824

Verma, B., Akinyi, M.V., Norppa, A.J., Frilander, M.J., 2018. Minor spliceosome and disease. Seminars in Cell & Developmental Biology 79, 103–112. https://doi.org/10.1016/j.semcdb.2017.09.036

Vieira, N.M., Naslavsky, M.S., Licinio, L., Kok, F., Schlesinger, D., Vainzof, M., Sanchez, N., Kitajima, J.P., Gal, L., Cavaçana, N., Serafini, P.R., Chuartzman, S., Vasquez, C., Mimbacas, A., Nigro, V., Pavanello, R.C., Schuldiner, M., Kunkel, L.M., Zatz, M., 2014. A defect in the RNA-processing protein HNRPDL causes limb-girdle muscular dystrophy 1G (LGMD1G). Human Molecular Genetics 23, 4103–4110. https://doi.org/10.1093/hmg/ddu127

Wahl, M.C., Will, C.L., Lührmann, R., 2009. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell 136, 701–718. https://doi.org/10.1016/j.cell.2009.02.009

Wan, Y., Larson, D.R., 2018. Splicing heterogeneity: separating signal from noise. Genome Biology 19, 86. https://doi.org/10.1186/s13059-018-1467-4

Wang, Y., Wu, X., Du, L., Zheng, J., Deng, S., Bi, X., Chen, Q., Xie, H., Férec, C., Cooper, D.N., Luo, Y., Fang, Q., Chen, J.-M., 2018. Identification of compound heterozygous variants in the noncoding RNU4ATAC gene in a Chinese family with two successive foetuses with severe microcephaly. Human Genomics 12, 3. https://doi.org/10.1186/s40246-018-0135-9

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J., Kristiansen, K., Krogh, A., Wang, J., Porse, B.T., 2012. Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biology 13, R35. https://doi.org/10.1186/gb-2012-13-5-r35

Will, C.L., Luhrmann, R., 2011. Spliceosome Structure and Function. Cold Spring Harbor Perspectives in Biology 3, a003707–a003707. https://doi.org/10.1101/cshperspect.a003707

Wong, J.J.-L., Au, A.Y.M., Ritchie, W., Rasko, J.E.J., 2016. Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. BioEssays 38, 41–49. https://doi.org/10.1002/bies.201500117

Wong, J.J.-L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., Thoeng, A., Khoo, T.-L., Bailey, C.G., Holst, J.,

Rasko, J.E.J., 2013. Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation. Cell 154, 583–595. https://doi.org/10.1016/j.cell.2013.06.052

Wu, Q., Krainer, A.R., 1996. U1-Mediated Exon Definition Interactions Between AT-AC and GT-AG Introns. Science 274, 1005–1008. https://doi.org/10.1126/science.274.5289.1005

Wu, Quiang, Krainer, A.R., 1999. AT-AC Pre-mRNA Splicing Mechanisms and Conservation ofMinor Introns in Voltage-Gated Ion Channel Genes. Molecular And Cellular Biology 19, 3225–3236.

Yeo, G.W., Nostrand, E.L.V., Liang, T.Y., 2007. Discovery and Analysis of Evolutionarily Conserved Intronic Splicing Regulatory Elements. PLoS Genetics 3, e85. https://doi.org/10.1371/journal.pgen.0030085

Zhang, C., Krainer, A.R., Zhang, M.Q., 2007. Evolutionary impact of limited splicing fidelity in mammalian genes. Trends in Genetics 23, 484–488. https://doi.org/10.1016/j.tig.2007.08.001