N° d'ordre NNT : 2016LYSE1280

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED340
Biologie Moléculaire, Intégrative et Cellulaire

Spécialité de doctorat : bioinformatique

Soutenue publiquement le 15/12/2016, par :
Clara BENOIT-PILVEN

Analyse de l'épissage alternatif dans les données RNAseq : développement et comparaison d'outils bioinformatiques

---

Devant le jury composé de :

VIEIRA-HEDDI Cristina, Professeure des universités, UCBL1	Présidente
BOEVA Valentina, Chargée de recherche, INSERM	Rapporteure
CORCOS Laurent, Directeur de recherche, INSERM	Rapporteur
LACROIX Vincent, Maitre de conférence, UCBL1	Examinateur
CHIKHI Rayan, Chargé de recherche, CNRS	Examinateur
AUBOEUF Didier, Directeur de Recherche, INSERM	Directeur de thèse

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
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
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. le Professeur Y. MATILLON
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

Remerciements

Je voudrais tout d'abord remercier les membres de mon jury : Cristina Vieira-Heddi, Valentina Boeva, Laurent Corcos, Vincent Lacroix et Rayan Chikhi d'avoir accepter d'assister à ma soutenance. Je souhaitais tout particulièrement remercier mes rapporteurs Valentina Boeva et Laurent Corcos d'avoir pris le temps de lire et évaluer mon manuscrit de thèse.

Je voudrais remercier mon directeur de thèse, Didier Auboeuf, de m'avoir accueilli dans son équipe pendant presque 4 ans, ainsi que tous les membres de l'équipe, anciens ou actuels : Amandine, Arnaud, Cyril, Emilie, Etienne, Fatima, Fabien, François-Olivier, Franck, Helen, Hélène, Hussein, Lamya, Léo, Lisa, Louis, Maira, Marie-Pierre, Sébastien, Simon et Sophie.

Je voudrais également remercier Vincent Lacroix et toute l'équipe Baobab-Erable pour m'avoir accueilli pour des réunions et des discussions, avec tant de gentillesse et d'attention.

Je remercie l'ARC santé 1 pour le financement sans lequel je n'aurai pas pu mener à bien mes travaux de doctorat.

Merci à mes amis les BIM, qu'ils soient restés sur Lyon ou pas, qu'ils aient partagés l'épreuve de la thèse ou pas : Amanda, Amandine, Anaïs, Béryl, Camille, Cindy, Hélène, Nada, Margaux, Matthieu, Ombeline et Vincent. Vous serez toujours ma famille Lyonnaise.

Merci aussi à mes amis Pictaviens : Aurore, Joséphine, Niels, Noé et Thibault.

Je remercie de tout mon coeur ma famille pour m'avoir soutenu dans les moments de doutes. Et surtout merci d'avoir toujours cru en moi.

Enfin, merci Vincent d'avoir toujours été là même quand j'étais insupportable et que je me plaignais tout le temps ! Merci de m'avoir supporté, encouragé, nourri !

Résumé

L'épissage alternatif est un processus biologique qui génère la diversité du protéome malgré le nombre limité de gène. Ce mécanisme régule à la fois les gènes de manières qualitatives (isoformes exprimées) mais aussi quantitatives (niveau d'expression). Avec le développement des technologies de séquençage à haut débit, il est maintenant possible d'étudier à large échelle les aspects quantitatifs et qualitatifs du transcriptome avec une même expérience (RNA-seq).

Durant ma thèse, j'ai développé une nouvelle méthode d'analyse de l'épissage alternatif dans les données RNA-seq. J'ai également participé à la mise en place du pipeline global d'analyse de données RNA-seq (expression et épissage). Cet outil a été utilisé pour analyser un grand nombre de jeux de données publics, générés par l'équipe et par des collaborateurs.

Dans un second temps, nous avons comparé notre outil d'analyse de l'épissage, FaRLine, qui est basé sur l'alignement sur un génome de référence, à KisSplice, une méthode basée sur l'assemblage. Nous avons montré que ces méthodes trouvaient un grand nombre d'événements en communs (70%), mais qu'il existait des différences non négligeables dues à la méthodologie. Nous avons analysé et classifié ces événements en 4 grandes catégories. Les variants faiblement exprimés et les exons chevauchant des éléments répétés sont mieux annotés par les méthodes basées sur l'alignement. Alors que les méthodes basées sur l'assemblage trouvent des nouveaux variants (exons ou sites d'épissage non annotés) et prédisent de l'épissage alternatif dans les familles de gènes paralogues. Notre travail souligne les points qui nécessitent encore l'amélioration des méthodes bioinformatiques.

Enfin, j'ai participé au développement de méthodes permettant d'aider les biologistes à évaluer l'impact fonctionnel de modifications d'épissage, que ce soit au niveau de la protéine produite en annotant les différents domaines protéiques au niveau des exons, ou à un niveau plus global en intégrant les modifications d'épissage dans les voies de signalisation.

Table des matières

1 Introduction **15**

I L'épissage . . . 17

A Mécanismes . . . 17

1 Réaction d'épissage . . . 18

2 Régulation de l'épissage . . . 19

3 Les différents types d'événements d'épissage . . . 22

B Importance de l'épissage alternatif . . . 24

C Épissage alternatif et pathologies . . . 27

II Les données à large échelle : le séquençage . . . 31

A Les technologies de séquençage haut-débit . . . 31

1 Les principes généraux du séquençage de seconde génération . . . 31

2 Les technologies de séquençage de seconde génération . . . 32

3 Les technologies de séquençage de troisième génération . . . 39

B Les différentes applications . . . 41

C L'analyse des données haut débit . . . 44

D Apports des données large échelle . . . 45

III Méthodes d'analyse de l'épissage . . . 47

A Historique de l'analyse de l'épissage : du cas par cas au large échelle . . . 47

1 La RT-PCR . . . 47

2 La PCR haut débit . . . 47

3 Les puces à ADN . . . 48

B Le séquençage du transcriptome : RNA-Seq . . . 50

C L'analyse de l'épissage dans les données RNA-seq . . . 51

1 Pipeline d'analyse de l'épissage alternatif . . . 52

2 Contrôle qualité . . . 52
3 Alignement . . . 54
4 Assemblage . . . 56
5 Identification des événements d'épissage ou des isoformes . . . 57
6 Quantification . . . 58
7 Analyse différentielle . . . 58
8 Visualisation . . . 59
9 Pour aller plus loin . . . 60

2 Objectifs **63**

3 Résultats **67**

I Développement d'un pipeline d'analyse de l'épissage alternatif dans les données RNA-Seq . . . 69
A FasterDB . . . 69
B FaRLine . . . 70
C Pipeline complet . . . 74
D Visualisation RNA-Seq . . . 75
E Validation de la méthode . . . 77
F Valorisation de la méthode . . . 78
G Discussion . . . 80
II Comparaison d'une méthode basée sur l'alignement à une basée sur l'assemblage 81
A Publication soumise à Genome Research . . . 81
B Discussion . . . 128
III Études de l'impact fonctionnelle de l'épissage alternatif . . . 129
A Annotation protéique à l'échelle des exons . . . 129
B Analyse des voies de signalisation . . . 131
C Discussion . . . 132

4 Conclusion et perspectives **135**

5 Annexes **139**

I Application à des données de collaborateurs . . . 141
II Exon ontologie . . . 156

Table des figures

1 Du pré-ARNm à la protéine 18
2 Sites de reconnaissances des exons et des introns 19
3 La réaction d'épissage 20
4 Régulation de l'épissage 21
5 Compléxité de la régulation de l'épissage 23
6 Évènements d'épissage 24
7 Évènement d'épissage complexe 24
8 Conséquences de l'épissage alternatifs 25
9 Epissage et NMD 26
10 Epissage et pathologies 28
11 Epissage et résistance 29
12 Principes généraux des technologies de séquençage de deuxième génération. . . . 32
13 Comparaison des différentes technologies de séquençage 33
14 L'amplification clonale des technologies NGS 34
15 Le séquençage 454 ou pyroséquençage 35
16 Le séquençage Illumina 36
17 Le séquençage SOLiD 37
18 Le séquençage Ion Torrent 38
19 Le séquençage PacBio 40
20 Le séquençage Oxford Nanopore 41
21 Les applications des NGS 42
22 Les formats fasta et fastq 45
23 La technique de RT-PCR 48
24 La puce à ADN 49

25 Pipeline classique d'analyse de l'épissage dans les données RNA-Seq 53
26 Analyse de l'épissage : la quantification 59
27 La visualisation des données RNA-Seq 61

28 Définition des exons génomiques dans FasterDB 70
29 Le pipeline d'analyse de l'épissage : FaRLine 71
30 Calcul du Ψ et du $\Delta\Psi$ pour un événement d'exon cassette 72
31 Redéfinition de certains types d'événement d'épissage 73
32 Le pipeline d'analyse de l'expression et de l'épissage 74
33 La visualisation RNA-Seq FasterDB 76
34 Validation expérimentale des événements de saut d'exon trouvés par FaRLine . . 78
35 Visualisation *Exon Ontologie* 130
36 L'intégration des données RNA-Seq dans les voies de signalisation 132

Liste des tableaux

1 Le séquençage haut-débit appliqué à l'épigénétique 43
2 Le séquençage haut-débit appliqué à la transcriptomique 44

Chapitre 1

Introduction

I L'épissage

L'expression des gènes est une suite de processus biologiques souvent considérés comme séquentiels. L'ADN présent dans le noyau est transcrit en ARN. Les ARN pré-messagers (pré-ARNm) produits sont maturés, puis exportés dans le cytoplasme et traduits en protéines.
L'étape de maturation des pré-ARNm en ARNs messagers matures est elle-même constituée de plusieurs étapes : ajout d'une coiffe en 5', polyadénylation en 3' et épissage. Cette dernière étape est un processus primordial qui permet de ne conserver que la partie codante des gènes.
Dans cette partie, nous verrons à quel point l'épissage est un mécanisme complexe et important dans la détermination du phénotype d'une cellule. Ensuite, nous aborderons le sujet des pathologies impliquant l'épissage alternatif.

A Mécanismes

Chez les organismes eucaryotes, la plupart des gènes sont morcelés. Ils sont composés d'une suite d'exons et d'introns. L'épissage est le processus qui consiste à exciser les introns du transcrit primaire et lier les exons entre eux pour former l'ARN messager mature (Figure 1). Il se déroule dans le noyau de la cellule au cours de la maturation des pré-ARNm grâce à l'intervention du spliceosome.

L'épissage qui consiste à garder tous les exons et à exciser tous les introns est appelé épissage constitutif. Mais il est très courant que certains introns soient conservés ou que certains exons soient éliminés (Figure 1). Ce processus est appelé épissage alternatif. Ce type d'altérations permet de former différentes protéines à partir d'un même gène ce qui explique en partie la grande diversité du protéome comparativement au nombre limité de gènes chez les organismes eucaryotes.

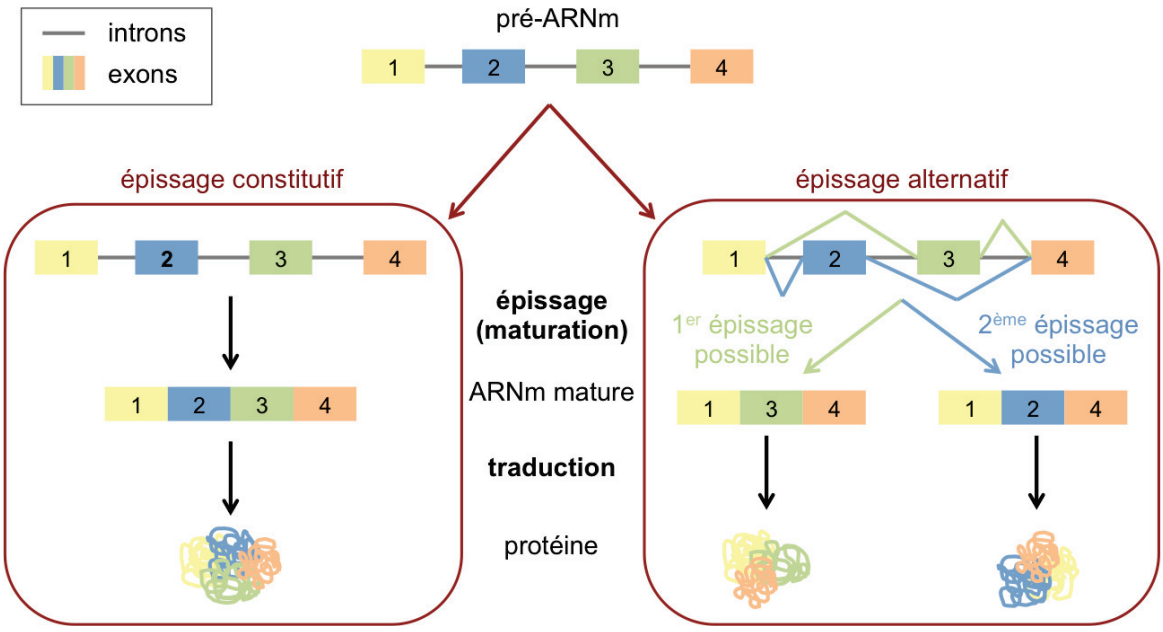


FIGURE 1 – **Du pré-ARNm à la protéine :** schéma explicatif simplifié de l'épissage constitutif et alternatif. L'épissage constitutif consiste en l'excision de tous les exons pour ne former qu'un ARNm mature contenant uniquement les exons. Cet ARNm peut être ensuite traduit en protéine. Dans le cas de l'épissage alternatif, des exons peuvent être inclus ou exclus. Un gène peut ainsi former différents ARNm matures et donc différentes protéines.

1 Réaction d'épissage

Afin d'obtenir des ARNs messagers fonctionnels, l'épissage doit être très spécifique et reproductible. Aussi, la reconnaissance des exons et des introns nécessite un ensemble de séquences caractéristiques et conservées [Mount, 1982]. Ces 3 sites sont (Figure 2 A) :

- le site 5', ou donneur, correspondant à la jonction exon-intron
- le site 3', ou accepteur, correspondant à la jonction intron-exon
- le site de branchement situé dans l'intron à environ 30 nucléotides en amont du site accepteur.

Dans une grande majorité des cas, les deux premières bases de l'intron situées au site 5' sont GU et les deux dernières bases situées au site 3' sont AG. À l'intérieur du site de branchement, on distingue une adénosine (A) qui joue un rôle clé pour la réaction d'épissage et est appelée point de branchement.

La similarité de ces séquences avec leur consensus définit la force des sites d'épissage [Yeo and Burge, 2004]. Plus la séquence d'un site est proche de la séquence consensus, plus le site est fort. Et inversement, plus elle est éloignée de la séquence consensus, plus le site sera faible (Figure 2 B). Cette force définie l'affinité de reconnaissance par les facteurs d'épissage. Des sites

forts induiront très généralement l'inclusion de l'exon (épissage constitutif). Alors que des sites faibles ne seront utilisés que dans certaines conditions physiologiques.

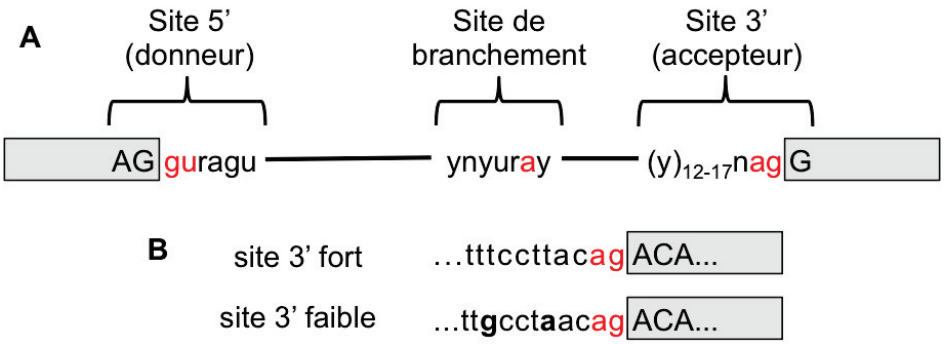


FIGURE 2 – **Sites de reconnaissances des exons et des introns. A.** Séquences consensus des sites d'épissage. Les nucléotides en rouge représentent les bases caractéristiques des séquences. **B.** Exemples de séquences pour deux sites accepteurs : un fort et un faible. Les nucléotides rouges indiquent le site 3' et les nucléotides en gras les bases ne correspondant pas au consensus. (Adapté de [Srebrow and Kornblihtt, 2006])

La reconnaissance de ces sites fait intervenir un grand complexe protéique, le spliceosome. La grande majorité des pré-ARNm est épissée via le spliceosome majeur (également appelé spliceosome U2-dépendant). Il est constitué de cinq petites ribonucléoprotéines nucléaires (snRNP) : U1, U2, U4/U6 et U5. Il comporte également des centaines d'autres protéines [Hegele et al., 2012, Wahl et al., 2009]. Près de 140 d'entre elles font partie du "core spliceosome", elles sont considérées comme les protéines centrales du complexe d'épissage. À travers plusieurs interactions protéine-protéine, protéine-ARN et ARN-ARN, le spliceosome reconnaît la jonction exon-intron et déclenche deux réactions de trans-estérification qui retirent l'intron sous la forme d'un lasso et lient les exons pour former le transcrit mature (Figure 3).

2 Régulation de l'épissage

La reconnaissance des sites d'épissage par le spliceosome est régulée par l'interaction de facteurs d'épissage (trans-régulateurs) avec des courtes séquences cis-régulatrices (∼10 nucléotides) du pré-ARNm. Par convention, ces éléments cis-régulateurs sont classés en fonction de leur place et en fonction de leur capacité à faciliter ou à empêcher l'épissage (Figure 4 A). Les séquences situées dans les exons sont nommées ESE (Exonic Splicing Enhancer) lorsqu'elles sont activatrices et ESS (Exonic Splicing Silencer) lorsqu'elles sont inhibitrices. Celles localisées dans les introns sont nommées ISE (Intronic Splicing Enhancer) et ISS (Intronic Splicing Silencer). Il existe de

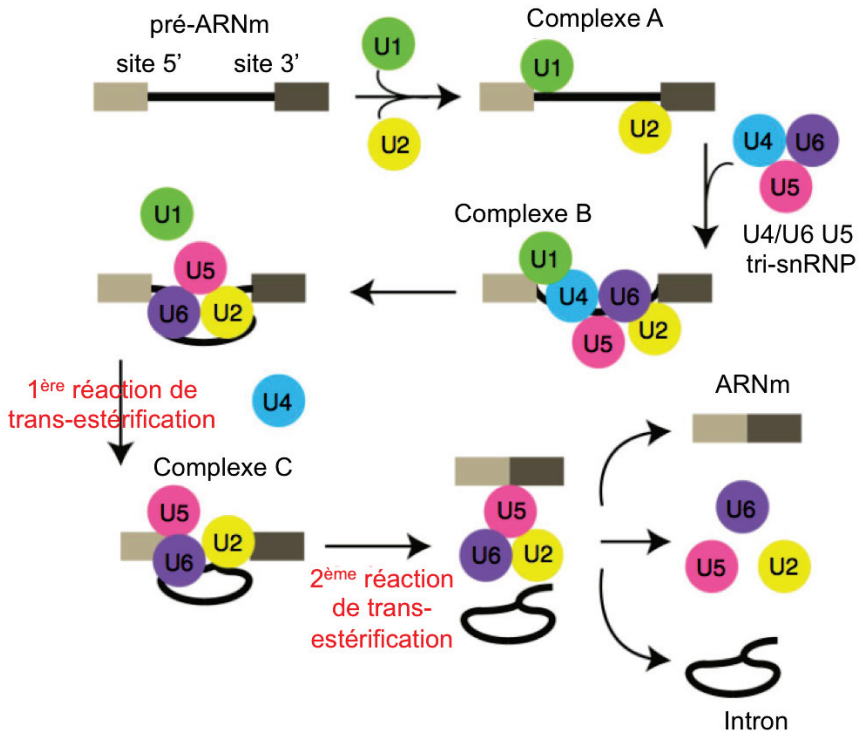


FIGURE 3 – **Schéma simplifié de la réaction d'épissage.** La première étape de l'assemblage du spliceosome consiste en la reconnaissance du site 5' par la snRNP U1. Puis la snRNP U2 vient se lier au site de branchement et forme le complexe A. Par la suite, le complexe B se forme grâce à la fixation du tri-snRNP U4/U6-U5 entre U1 et U2. Après la libération des snRNPs U1 et U4, la première réaction de trans-estérification a lieu. Le site 5' est clivé et le résidu en 5' de l'intron est lié à une séquence proche du site 3' (complexe C). La deuxième réaction de trans-estérification consiste à cliver le site 3', produisant les exons liés entre eux et l'intron sous forme de lasso (appelé lariat). (Adapté de [Yoshida and Ogawa, 2014])

nombreux facteurs d'épissage. Les plus connus sont les ribonucléoprotéines hétérogènes nucléaires (hnRNP) pour leur rôle inhibiteur et les protéines SR pour leur rôle plutôt activateur (Figure 4 B).

Le positionnement des séquences cis-régulatrices et l'abondance relative des facteurs d'épissage ne permettent pas d'expliquer la fine régulation de l'épissage. Le couplage entre la transcription et l'épissage, la conformation de la chromatine ou encore les structures secondaires des ARNs sont d'autres éléments qui jouent un rôle dans le choix des exons qui vont être inclus dans les transcrits matures.

La transcription et la maturation des ARN ont longtemps été considérées comme des processus séquentiels. Or ces deux mécanismes peuvent avoir lieu en même temps et s'influencer l'un l'autre [Bentley, 2002, Proudfoot and O'Sullivan, 2002]. Notamment, l'épissage consti-

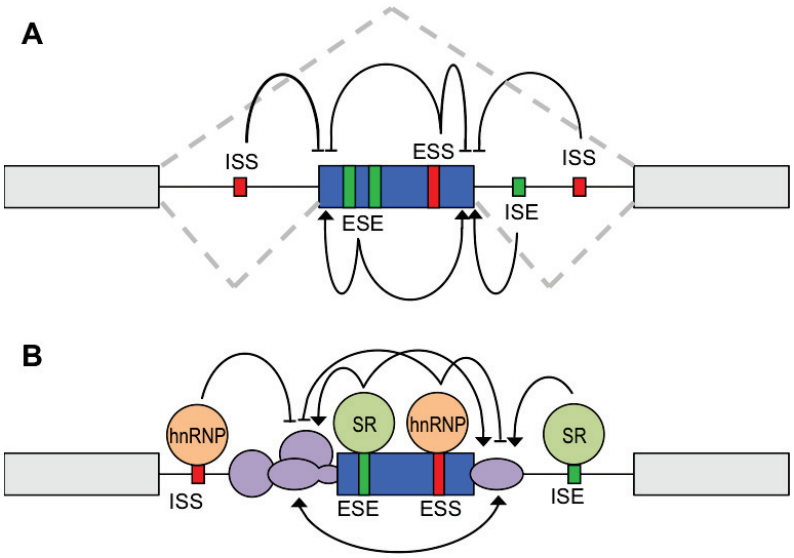


FIGURE 4 – **Régulation de l'épissage. A.** Les séquences cis-régulatrices. Les séquences sont nommées d'après leur position et leur activité. (Adapté de [Matlin et al., 2005]) **B.** Les facteurs d'épissage (trans-régulateurs) interagissent avec les séquences cis-régulatrices pour promouvoir l'inclusion ou l'exclusion de l'exon. Les deux familles de facteurs de régulation de l'épissage les plus connues sont les protéines hnRNP (en orange) et les protéines SR (en vert). En violet sont représentés des composants du spliceosome. (Adapté de [Kornblihtt et al., 2013])

tutif semble essentiellement co-transcriptionnel alors que l'épissage alternatif ne l'est pas toujours [Vargas et al., 2011]. Deux mécanismes, non mutuellement exclusifs, expliquent l'influence de la transcription sur l'épissage. Tout d'abord, la machinerie transcriptionnelle est capable de recruter des facteurs d'épissage qui vont permettre de réguler l'inclusion ou l'exclusion des exons. Ce recrutement est notamment réalisé par le CTD (Carboxy-Terminal Domain) de l'ARN polymérase II. Ce CTD ne sert pas uniquement à recruter des protéines mais il régule aussi leur activité. Ensuite, la cinétique d'élongation de l'ARN polymérase II joue un rôle dans la régulation de l'épissage. En effet, plus l'élongation va être rapide, moins les exons avec des sites d'épissage faibles vont être reconnus et donc inclus. Une grande vitesse d'élongation induirait donc l'exclusion alors qu'une faible vitesse d'élongation, l'inclusion de ces exons. Cependant, il y a des exceptions : dans certaines conditions particulières, une faible vitesse d'élongation a été montrée comme favorisant l'exclusion d'exons [Ip et al., 2011, Dutertre et al., 2010].

Un autre paramètre entre en jeu dans la régulation de l'épissage alternatif, c'est la conformation de la chromatine. De nouveau, deux points importants ont été identifiés : les modifications d'histones et le positionnement des nucléosomes. Les histones sont les protéines qui s'associent à l'ADN pur former la structure de base de la chromatine. Elles jouent un rôle important dans l'empaquetage et le repliement de l'ADN. Ces protéines peuvent subir des modifications post-

traductionnelles, comme la méthylation, l'acéthylation ou la phosphorylation. Ces modifications d'histones permettent de moduler la structure de la chromatine et influencent ainsi des processus cellulaires tels que la transcription. Les modifications d'histones régulent l'épissage alternatif en activant ou en inhibant la transcription. Or comme la cinétique de transcription permet de réguler la reconnaissance des sites d'épissage, un changement de modification d'histone modifie la cinétique de transcription et donc l'épissage. De plus, il a été montré que les modifications d'histones régulent l'épissage via le recrutement de facteurs d'épissage [Luco et al., 2010]. Le positionnement des nucléosomes semblent également être un facteur régulateur. En effet, de nombreuses études ont montré que les nucléosomes se positionnaient préférentiellement au niveau des exons [Schwartz et al., 2009, Spies et al., 2009, Tilgner et al., 2009, Iannone et al., 2015]. De plus, ce positionnement des nucléosomes au niveau des exons est plus courant pour les exons avec des sites d'épissage faibles [Spies et al., 2009, Tilgner et al., 2009], comme pour signifier à la machinerie d'épissage qu'un exon est présent.

La formation de structure secondaire par les ARNs est un autre facteur pouvant réguler l'épissage alternatif. En effet, ce type de repliement local des molécules d'ARN peut influer sur l'épissage de plusieurs façons. En masquant des sites d'épissage, le spliceosome ne pourra pas reconnaître le site et donc l'épissage de l'exon en question sera inhibé. En occultant ou en exposant des séquences cis-régulatrices, le repliement de l'ARN peut promouvoir ou empêcher un épissage. L'appariement entre une séquence intronique en amont et une autre en aval d'un exon forme une boucle qui inclut cet exon et entraîne ainsi son exclusion.

La quantité de paramètres entrant en jeu dans la régulation de l'épissage (Figure 5) rend l'objectif de définir un "code de l'épissage" très difficile.

3 Les différents types d'événements d'épissage

La régulation de l'épissage étant très complexe et très fine, le choix des sites d'épissage donnent lieu à plusieurs types d'événements d'épissage (Figure 6). On appelle événement d'épissage une portion de gène qui peut produire plusieurs variants. Un événement est composé de deux variants possibles. L'événement le plus courant est le saut d'exon, encore appelé exon cassette : l'exon est inclus ou exclu de l'ARN mature. Il arrive que plusieurs exons soient exclus ou inclus en même temps, on parle alors de saut d'exons multiple. Deux exons consécutifs sont dits mutuellement exclusifs lorsque l'inclusion d'un exon entraîne l'exclusion de l'autre. Enfin, il n'y a pas que les exons complets qui peuvent être alternativement épissés. Des morceaux d'exons peuvent être

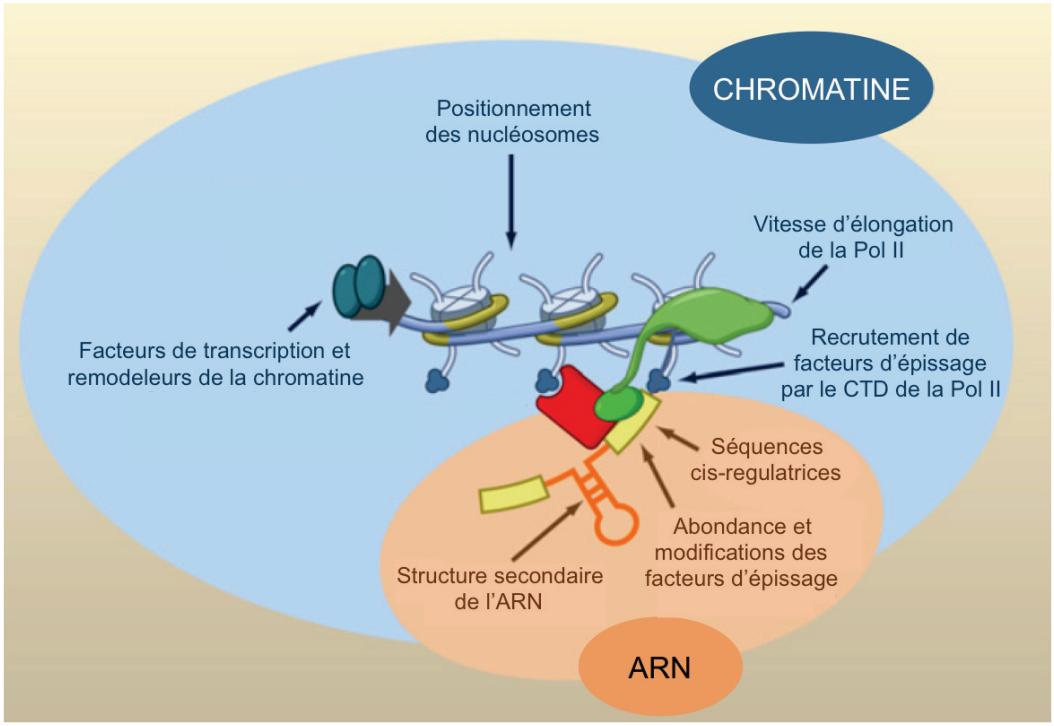


FIGURE 5 – **Complexité de la régulation de l'épissage.** L'épissage alternatif est le résultat de la combinaison de nombreux paramètres : les éléments cis-régulateurs et les structures secondaires de l'ARN, mais aussi les propriétés de la transcription et de la conformation de la chromatine qui régulent le recrutement des facteurs d'épissage sur le pré-ARNm.(Adaptée de [Luco et al., 2010])

inclus ou exclus par le choix d'un site donneur (en 5') ou d'un site accepteur (en 3') alternatif. Des introns peuvent également être inclus dans le transcrit mature. Ce type d'événement est nommé rétention d'intron.

D'autres types de modifications permettent de produire différents variants à partir d'un même gène, mais ils ne sont pas à proprement parler dus à l'épissage. En effet, une modification des sites d'initiation ou de terminaison de la transcription peut entraîner l'inclusion d'un premier ou dernier exon différent. Ces types d'événements sont appelés promoteurs alternatifs et sites de polyadélynation alternatifs. Ils sont très souvent analysés avec les événements d'épissage.

Cette séparation en type d'événements d'épissage n'est en réalité pas aussi simple. La définition d'un événement d'épissage comme une comparaison de deux variants d'épissage limite nos analyses. En effet, très souvent plusieurs types d'événements peuvent être couplés et former des événements complexes. La figure 7 montre l'exemple d'un événement couplant un site accepteur alternatif avec le saut d'un exon. À partir de ce gène, 4 variants différents peuvent potentiellement être générés. Afin d'étudier correctement ce type d'événement complexe, il faut revoir notre manière de définir les événements d'épissage.

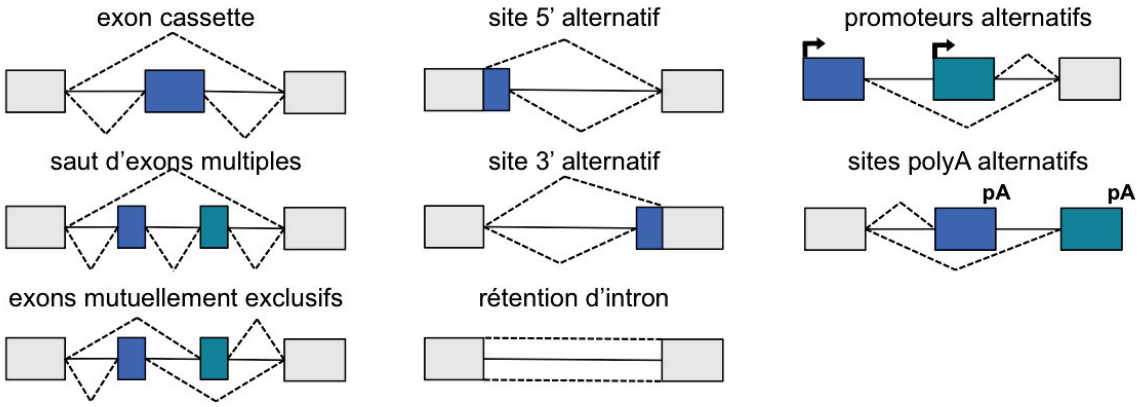


FIGURE 6 – **Différents événements produisant plusieurs variants à partir d'un même gène.** Les cinq premiers types sont directement dus à l'épissage alternatif. Les deux derniers proviennent d'un changement d'initiation ou de terminaison de la transcription.

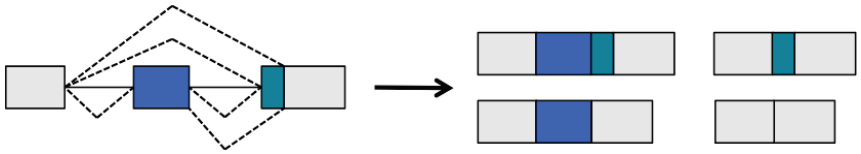


FIGURE 7 – **Évènement d'épissage complexe**, couplant à la fois un site accepteur alternatif et un saut d'exon. Ce type d'événement peut produire quatre variants différents.

B Importance de l'épissage alternatif

A sa découverte, l'épissage alternatif était considéré comme marginal [Berget et al., 1977] [Chow et al., 1977]. Or, parmi les 90% des gènes humains annotés comme multi-exoniques [Cusack et al., 2011], on estime maintenant qu'environ 95% subissent de l'épissage alternatif [Pan et al., 2008, Barash et al., 2010]. L'épissage alternatif est donc une règle et non pas une exception.

Une grande proportion des épissages alternatifs touche la région codante des ARNs, permettant parfois de produire plusieurs protéines à partir d'un même gène. L'épissage alternatif permet ainsi d'expliquer la diversité du protéome. De plus, les protéines produites à partir d'un même gène peuvent avoir des fonctions différentes voire opposées.

Par exemple, le gène FAS peut produire un variant avec une fonction pro-apoptotique (induction de la mort de la cellule appelée apoptose) et un autre variant avec une fonction anti-apoptotique [Izquierdo et al., 2005]. Le variant d'épissage du gène FAS incluant tous les exons produit une protéine transmembranaire. Lorsque le ligand de FAS, nommé FASL, se lie au récepteur, le signal pour déclencher l'apoptose est transmis aux effecteurs. La cellule va alors se détruire. Mais lorsque l'exon 6 du gène FAS est exclu lors de l'étape de maturation, la protéine

produite est soluble. Lors de l'activation du recepteur par son ligand, le signal ne peut pas être transmis aux effecteurs. L'apoptose est alors inhibée (Figure 8 A).

Le gène BLC-X peut également produire un variant pro-apoptotique et un autre anti-apoptotique [Xerri et al., 1998]. La seule différence entre ces deux variants est le choix du site donneur de l'exon 2 (Figure 8 B).

Un dernier exemple est le facteur de croissance des cellules endothéliales vasculaires (VEGF). Ce gène génère un grand nombre de variants dont certaines formes vont favoriser la formation de nouveaux vaisseaux sanguins et d'autres non [Bates et al., 2002, Guyot and Pagès, 2015]. La différence entre les variants pro-angiogéniques (qui favorise la formation de nouveaux vaisseaux sanguins) et les variants anti-angiogéniques se trouve dans le choix du dernier exon. Si l'exon 8a est inclus, alors le 8b est exclu et la protéine produite est pro-angiogénique. A l'inverse, si l'exon 8b est inclus, le 8a est exclu et la protéine produite est anti-angiogénique (Figure 8 C).

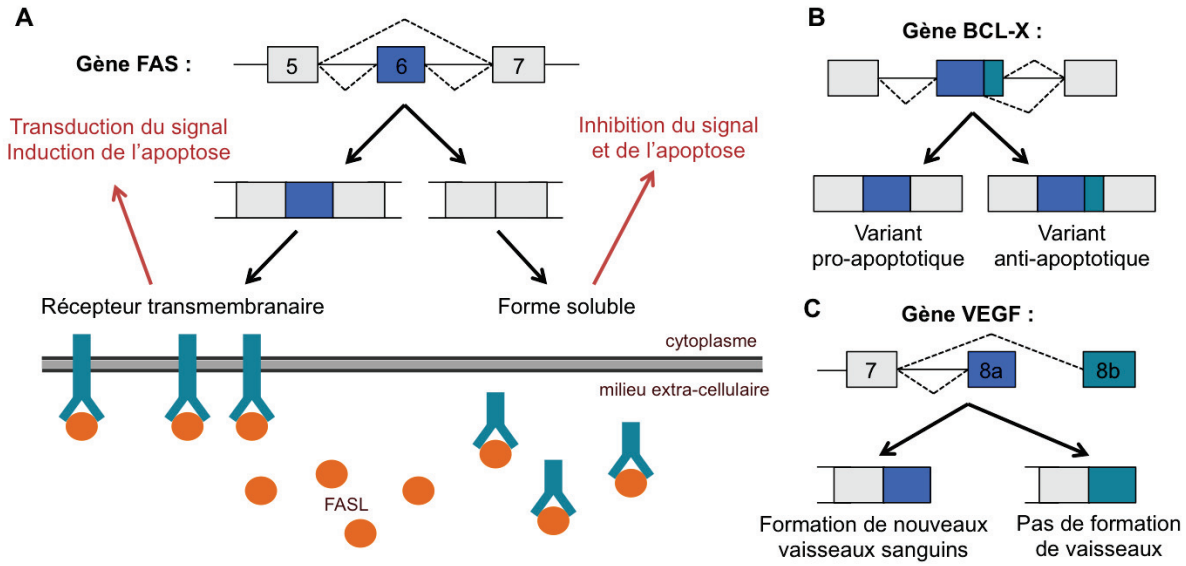


FIGURE 8 – **Conséquences de l'épissage alternatifs.** Quelques exemples d'épissage alternatif donnant lieu à des protéines avec des fonctions très différentes. **A.** L'exclusion de l'exon 6 lors de l'épissage du gène FAS produit des variants ayant une fonction anti-apoptotique. Alors que les variants incluant l'exon 6 ont une fonction pro-apoptotique. **B.** Le changement du site donneur dans l'exon 2 du gène BCL-X conduit à la production de variants avec des fonctions opposées. L'inclusion de la forme courte de l'exon 2 donne un variant pro-apoptotique. Alors que l'inclusion de la forme longue de l'exon 2 donne une forme anti-apoptotique. **C.** Le choix du dernier exon lors de l'épissage du gène VEGF confère aux protéines produites soit une fonction pro-angiogénique, soit une fonction anti-angiogénique.

En plus de cette régulation qualitative, l'épissage permet de modifier quantitativement l'expression des gènes, en produisant des transcrits non fonctionnels. Ce mécanisme est nommé RUST

pour "Regulated Unproductive Splicing and Translation". En effet, une modification d'épissage peut potentiellement introduire un codon stop prématuré (PTC, Premature Stop Codon). Le transcrit est alors pris en charge par la machinerie de NMD (Nonsense-Mediated mRNA Decay) pour être dégradé [Lewis et al., 2003, Weischenfeldt et al., 2012]. De tels transcrits peuvent provenir de différents types d'événements d'épissage : soit par l'ajout d'un exon/intron contenant un codon de terminaison dans le cadre de lecture ou par l'ajout d'un exon/intron venant modifier le cadre de lecture et ainsi introduire un PTC en aval.

Ce mécanisme permet également de dégrader des variants "anormaux", comme pour le facteur de croissance FGFR2. Ce gène contient deux exons mutuellement exclusifs. Le mécanisme de NMD dégrade les variants contenant les deux exons ou aucun des deux. Alors que les variants contenant un seul de ces exons produisent des protéines (Figure 9).

Le NMD est le plus souvent présenté comme un mécanisme de contrôle qualité des ARNs messagers matures. Son rôle dans le RUST n'a été découvert que plus récemment [Lareau et al., 2007]. Il est intéressant de noter que de nombreux facteurs d'épissage s'auto-régulent grâce à ce mécanisme, comme PTB ou SC35 [Lareau et al., 2007].

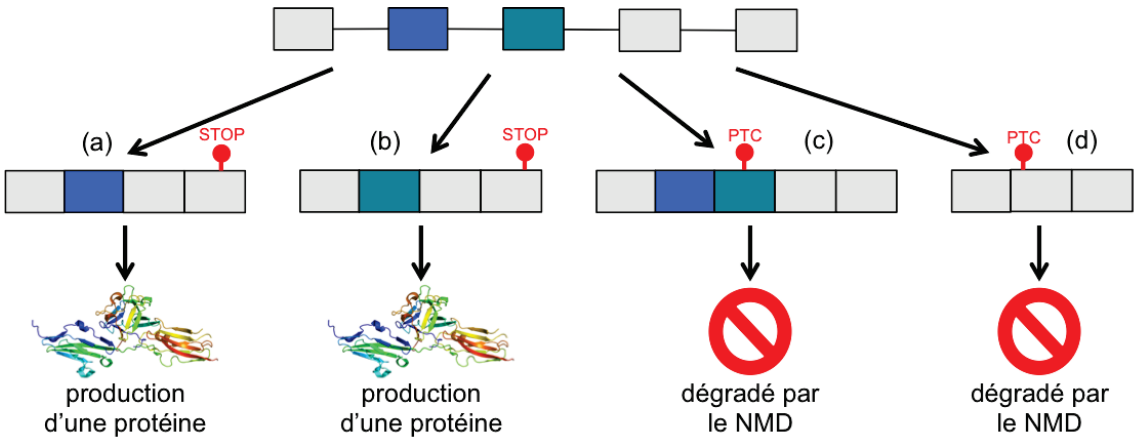


FIGURE 9 – **Epissage et NMD.** Le facteur de croissance FGFR2 produit plusieurs variants d'épissage. L'inclusion des deux exons mutuellement exclusifs (c) ou aucun des deux (d) entraîne la dégradation des variants par le mécanisme de NMD à cause de l'apparition d'un codon stop prématuré. Au contraire, les variants contenant uniquement un des deux exons mutuellement exclusifs (a et b) produisent des protéines fonctionnelles. (Adapté de [Lareau et al., 2007])

C Épissage alternatif et pathologies

De par sa capacité à modifier qualitativement et quantitativement les transcrits et donc les protéines produites, l'épissage alternatif est impliqué dans de nombreuses maladies. Les altérations de l'épissage impliquées dans les maladies peuvent être dues à des mutations dans des séquences cis-régulatrices, dans des facteurs d'épissage ou encore dans des éléments majeurs du spliceosome. En effet, on estime aujourd'hui qu'environ un tiers des maladies causées par des mutations d'un seul nucléotide (SNP) entraîne une modification de l'épissage [Singh and Cooper, 2012].
Une mutation dans une séquence cis-régulatrice peut avoir différentes conséquences. La première est la modification de l'événement d'épissage : exclusion d'un exon constitutif ou rétention d'un intron (Figure 10 A). Sans changer totalement l'événement d'épissage, une mutation peut également modifier le ratio d'inclusion d'un exon (Figure 10 B). Par exemple, la démence frontotemporale avec Parkinsonisme implique une mutation dans l'exon 10 du gène MAPT qui modifie le ratio entre les 2 isoformes de la protéine produite. Enfin, une mutation ponctuelle peut aussi créer un site d'épissage cryptique entraînant l'inclusion d'un pseudo-exon ou l'utilisation d'un nouveau site donneur ou accepteur (Figure 10 C). La mucoviscidose [Friedman et al., 1999], l'ataxie télangiectasie [Du et al., 2007], la neurofibromatose de type 1 [Pros et al., 2009] sont des exemples de maladies parmi de très nombreuses autres [Pérez et al., 2010] impliquant l'inclusion de pseudo-exons.

Des maladies peuvent également impliquer des modifications de régulateurs de l'épissage. Par exemple, la Dystrophie Myotonique de type 1 (DM1) trouve sa cause dans une expansion anormale du tri-nucléotide CTG dans la région 3'UTR du gène DMPK [Lee and Cooper, 2009]. La pathogénicité de ces répétions est due à un gain de fonction des ARNm produits. Ils forment des agrégats nucléaires qui séquestrent les facteurs de transcription MBNL1 et CUGBP1. Cette séquestration entraîne une perte de fonction de ces facteurs d'épissage causant de nombreux défauts d'épissage habituellement régulés par ces facteurs.
Enfin, de nombreuses maladies dues à des mutations qui affectent directement la machinerie d'épissage ont été décrites. On trouve par exemple la rétinite pigmentaire dont la forme autosomique dominante est associée à une mutation dans les gènes PRPF3, PRPF8, PRPF31 et SNRNP200. Ces gènes codent pour des protéines importantes dans l'assemblage du spliceosome majeur, notamment l'assemblage de U4/U6 avec U5. L'amyotrophie spinale est un autre exemple de maladie associée à des mutations affectant la machinerie d'épissage. Cette maladie est due

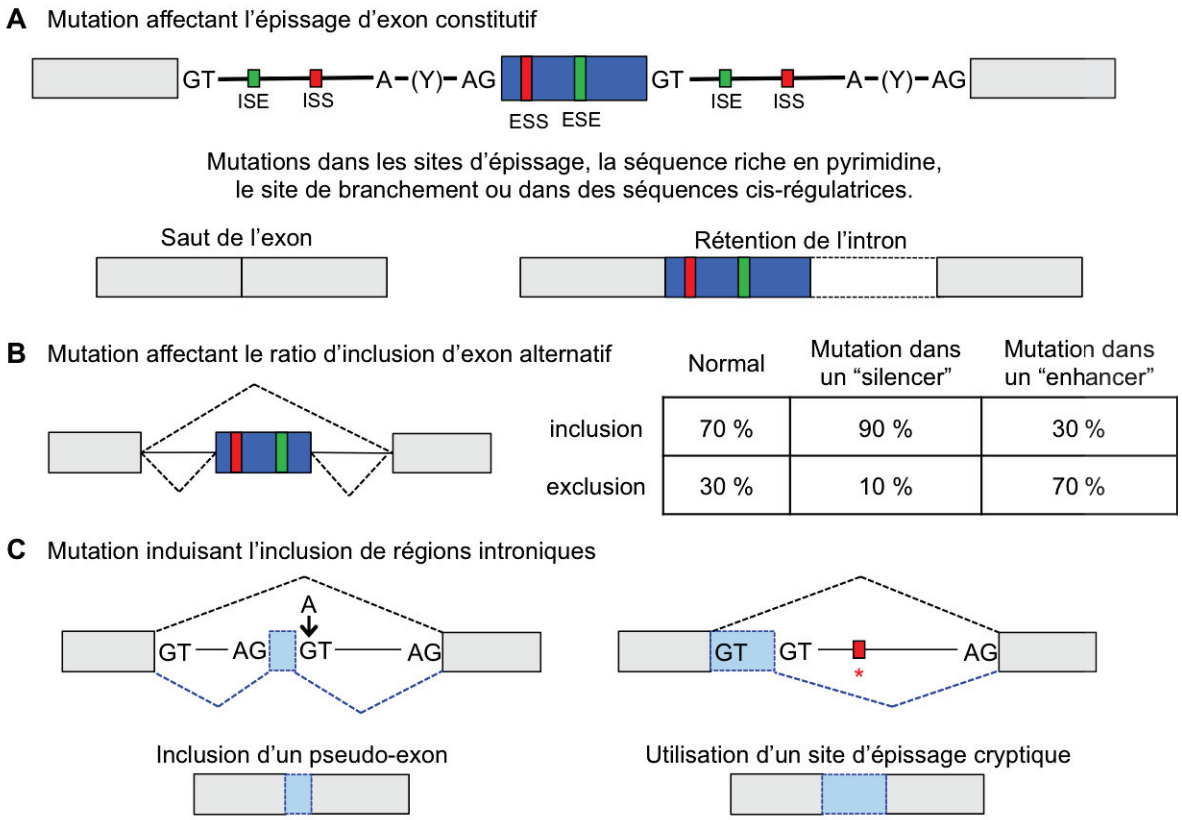


FIGURE 10 – **Epissage et pathologies.** Trois mécanismes par lesquels une mutation affectant l'épissage peut agir. **A.** Une mutation ponctuelle dans un site d'épissage, le site de branchement, la séquence riche en pyrimidine ou encore les éléments cis-régulateurs peut entraîner le saut d'un exon ou l'inclusion d'un intron. **B.** Les mutations dans les séquences cis-régulatrices peuvent également modifier le ratio d'inclusion d'un exon alternatif. Une mutation dans une séquence "silencer" entraînera l'augmentation de l'inclusion de l'exon. Alors qu'une mutation dans une séquence "enhancer" entraînera une diminution de l'inclusion de l'exon. **C.** Des mutations dans les introns peuvent mener à l'inclusion d'une région intronique (représentée en bleu clair). La reconnaissance du site d'épissage cryptique est due à une mutation créant soit un site d'épissage (indiqué par la flèche), soit un site "enhancer" (indiqué par l'astérisque rouge). (Adapté de [Singh and Cooper, 2012])

à la perte du produit du gène SMN1. Or les gènes de la famille SMN sont indispensables pour l'assemblage du spliceosome [Terns and Terns, 2001]. Enfin, des mutations directement dans les snRNP du complexe d'épissage provoquent des maladies, comme le syndrome de Taybi-Linder. Cette pathologie est causée par des mutations dans le snRNP U4-atac qui est un composant du spliceosome mineur [Edery et al., 2011, He et al., 2011].

Il existe de nombreuses autres maladies qui ont été associées à une dérégulation de l'épissage alternatif [Wang and Cooper, 2007, Tazi et al., 2009], dont notamment le cancer. En effet, on trouve un très grand nombre de modifications d'épissage dans les cancers. L'expression de variants d'épissage spécifiques aux tumeurs affecte de nombreux processus cellulaires critiques dans

la biologie du cancer, comme la prolifération, la motilité ou l'apoptose. L'épissage a notamment été montré comme étant impliqué dans la progression tumorale et dans la formation de métastases. Voici trois exemples de gènes dont l'épissage a été montré comme jouant un rôle dans certains cancers.

Il a été montré que la perte de l'expression du gène SYK est associée à une invasivité accrue dans certains cancers du sein [Coopman et al., 2000, Toyama et al., 2003]. Mais l'expression de ce gène n'est pas le seul facteur jouant un rôle important dans certains cancers du sein. En effet, ce gène peut exprimer plusieurs isoformes. L'isoforme longue possède une activité de suppresseur de tumeur [Wang et al., 2003]. Alors que la forme la plus courte qui est exprimée uniquement dans les cellules cancéreuses, n'empêche pas la progression tumorale et la formation de métastases. Ce n'est donc pas simplement la perte de l'expression du gène SYK qui est en cause dans certains cancers du sein mais la perte de l'expression de l'isoforme longue de ce gène.

Le gène codant pour la glycoprotéine transmembranaire CD44 peut produire un grand nombre de variants par épissage alternatif. Ce gène exprime de très nombreux variants dont certains sont impliqués dans la progression tumorale et la formation de métastases [Prochazka et al., 2014, Orian-Rousseau, 2015].

Enfin, il a été montré que l'épissage du gène CCND1 impactait la formation de métastases dans le cancer de la prostate [Lapuk et al., 2014].

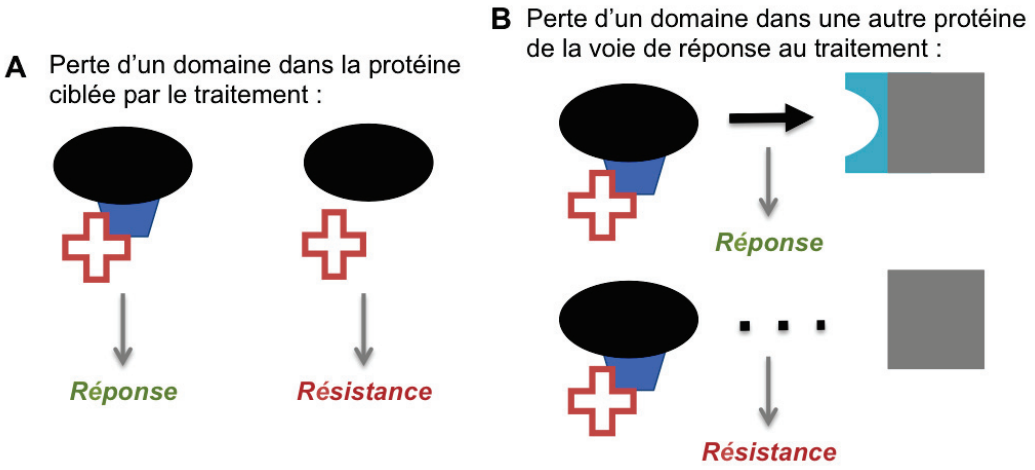


FIGURE 11 – **Epissage et résistance.** Mécanismes par lesquels l'épissage peut entraîner une résistance à une thérapie ciblée. **A.** La protéine ciblée par le traitement peut être directement affectée par une modification d'épissage. Par exemple, le domaine d'interaction avec la molécule thérapeutique est perdu. **B.** L'épissage peut également affecter d'autres gènes de la voie de réponse au traitement. Si un domaine d'interaction n'est plus fonctionnel, la voie de signalisation sera affectée et donc la réponse au traitement aussi.

L'épissage est également impliqué dans la réponse aux traitements. Ces dernières années, de plus en plus de cas de résistance à des thérapies anticancéreuses dues à l'épissage alternatif ont été décrits. La formation de protéines tronquées ou non fonctionnelles à n'importe quel niveau de la voie de réponse au traitement peut potentiellement entraîner une résistance (Figure 11).

II Les données à large échelle : le séquençage

Le but du séquençage est de déterminer la suite des nucléotides qui compose une séquence d'ADN. En 1977, le séquençage Sanger est développé [Sanger et al., 1977]. Cette méthode permet de déterminer avec précision la séquence d'intérêt, mais ne permet d'analyser qu'une seule séquence à la fois. Avec l'automatisation de cette approche et la commercialisation en 2002 du 3730 DNA Analyzer de Applied Biosystems (aujourd'hui Life Technologies), on voit apparaître la première génération de séquençage haut-débit. C'est avec cette technologie que la première ébauche du génome humain fut terminée en 2004.

En 2005, apparaît la deuxième génération de séquençage haut-débit ou "Next Generation Sequencing" (NGS) avec la technologie 454 Life Science (aujourd'hui, Roche). Les NGS ont grandement accéléré la recherche en biologie et biomédecine en rendant l'analyse du génome et du transcriptome beaucoup plus rapide et moins coûteuse. Aujourd'hui, l'utilisation du séquençage haut-débit est largement répandue et est même devenue la norme pour les analyses à large échelle. Mais le développement de ces technologies a nécessité et nécessite toujours des progrès au niveau bioinformatique, afin d'obtenir des algorithmes d'analyse performants et fiables, adaptés à chaque utilisation.

A Les technologies de séquençage haut-débit

1 Les principes généraux du séquençage de seconde génération

Plusieurs étapes sont communes aux différentes technologies de séquençage (Figure 12). La première est la création de la banque d'ADN simple brin associé à des ligands. Il est important de noter que pour des analyses sur le transcriptome, l'ARN n'est pas séquencé directement, mais il est converti par retrotranscription en ADN complémentaire (ADNc). La préparation de cette banque d'ADN que l'on appelle librairie, nécessite une fragmentation pour obtenir des séquences de la taille souhaitée. Puis, des adaptateurs sont ajoutés aux extrémités des fragments obtenus. Ils vont permettre la fixation des fragments (sur des billes ou sur une plaque) et leur amplification. Ils servent également à identifier les fragments appartenant à un échantillon lorsque l'on séquence plusieurs échantillons en même temps. Ce processus appelé multiplexage est rendu possible par l'ajout de "code-barre" (suite nucléotidique spécifique de chaque échantillon) dans les adaptateurs.

Lors de la deuxième étape, la librairie obtenue est amplifiée afin d'avoir suffisamment de matériel

pour le séquençage. La méthode d'amplification varie suivant les technologies de séquençage. Enfin, vient l'étape du séquençage qui dépend elle aussi de la technologie utilisée. C'est au cours de cette étape qu'est déterminée la suite de nucléotide des fragments.

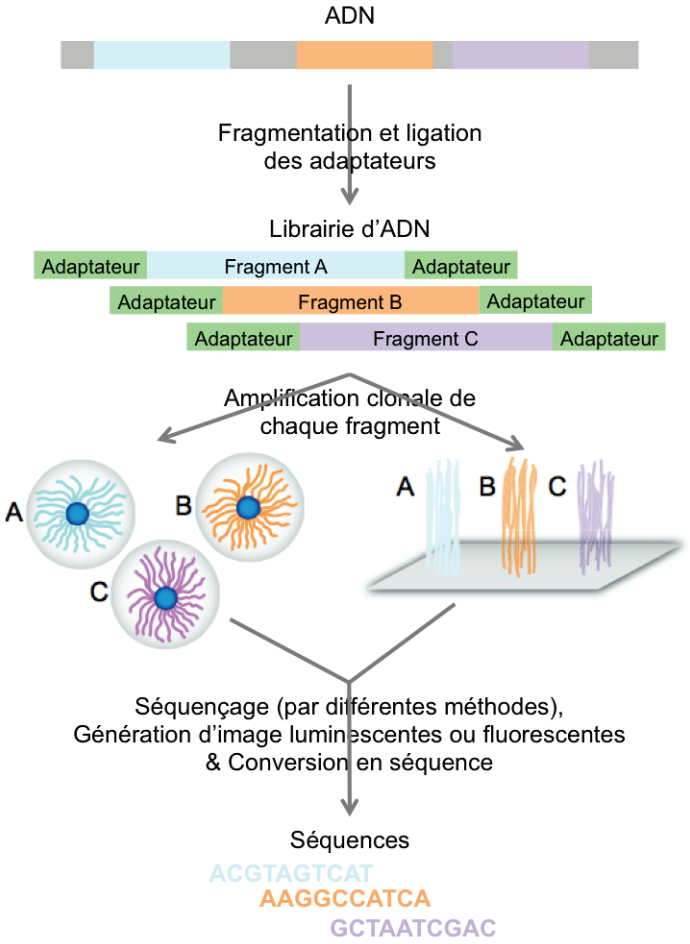


FIGURE 12 – **Principes généraux des technologies de séquençage de deuxième génération.** Les 3 étapes sont la préparation de la librairie, l'amplification de celle-ci et le séquençage. L'amplification peut être réalisée soit par une PCR en émulsion (schéma de gauche) soit par la formation de pont sur une surface (schéma de droite).

2 Les technologies de séquençage de seconde génération

Dans cette partie, nous présenterons les 4 technologies NGS qui ont dominé le marché durant la dernière décennie : 454, Illumina, SOLiD et Ion Torrent. Ces technologies ont toutes des spécificités quant à la longueur des lectures et la profondeur de séquençage (Figure 13).

454

Le séquenceur 454 fut le premier à être commercialisé en 2005. Cette technologie permet de séquencer des lectures de 400 nucléotides en moyenne (1000 nucléotides au maximum) à une

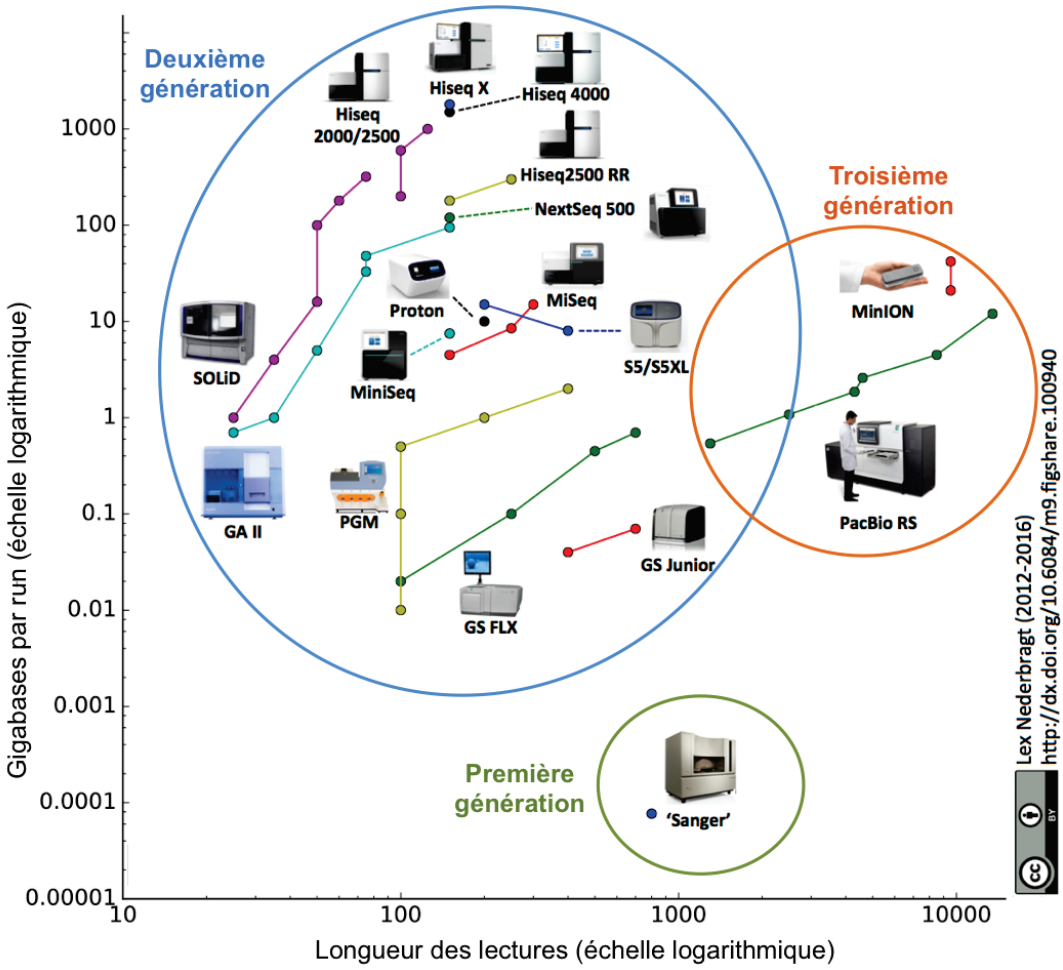


FIGURE 13 – **Comparaison des différentes technologies de séquençage.** Graphique représentant la profondeur de séquençage en fonction de la longueur des lectures pour les différentes technologies de séquencage de première, deuxième et troisième génération. (Adaptée d'une figure créée par Lex Nederbragt `http://dx.doi.org/10.6084/m9.figshare.100940`)

profondeur variant de 200 000 à 1 000 000 de lectures suivant la machine utilisée.

L'amplification clonale des fragments d'ADN est réalisée par PCR en émulsion (Figure 14 A). Les fragments sont d'abord fixés sur des billes (un fragment par bille). Puis, les billes sont introduites dans une émulsion contenant les réactifs nécessaires à la PCR. Chaque micro-goutte de l'émulsion va contenir une seule bille et ainsi permettre l'amplification de chaque fragment individuellement. Les billes sont par la suite distribuées dans plusieurs millions de puits dont le diamètre va permettre de ne récupérer qu'une seule bille par puits.

Le séquençage des fragments est ensuite réalisé par la technique de pyroséquençage (Figure 15). Chaque puits contient les enzymes et primers nécessaires à la synthèse d'ADN sauf les nucléotides. Les dNTP (A, T, G, C) sont ajoutés un par un séquentiellement et un lavage est effectué entre

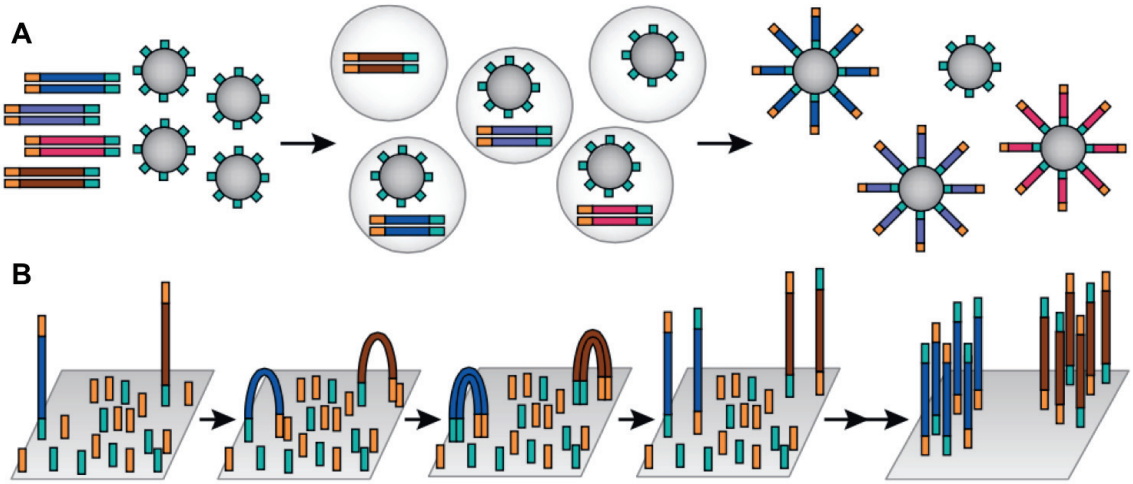


FIGURE 14 – **Les deux méthodes d'amplification clonale utilisées par les technologies de séquençage de seconde génération. A.** L'amplification par PCR en émulsion est utilisée par la technologie 454, SOLiD et Ion Torrent. Les fragments d'ADN sont fixés sur une bille et l'amplification est réalisée dans des gouttelettes qui contiennent tous les réactifs nécessaires à la PCR. **B.** La méthode d'amplification par formation de pont est celle utilisée pour le séquençage Illumina. Les fragments fixés sur la surface vont être amplifiés par PCR en formant des ponts avec les adaptateurs recouvrants la surface. À la fin des cycles de PCR, chaque cluster contient environ 1 000 copies du fragment d'ADN d'origine. [Shendure and Ji, 2008]

chaque introduction de nucléotide. Les nucléotides vont être incorporés pour synthétiser le brin d'ADN complémentaire. Lorsqu'un nucléotide est incorporé, un signal lumineux est émis. Il est capté par un système de détection de luminescence. Les images enregistrées sont ensuite traitées informatiquement pour être traduites en séquences.

Grâce à ses longues lectures, cette technologie fut pendant un moment la méthode la plus utilisée pour l'assemblage *de novo* de génome et la métagénomique. Mais, cette technologie a des désavantages comme la faible profondeur de séquençage et le fort taux d'erreur dans les répétitions d'homopolymères. Un homopolymère est une séquence répétant la même base (AAAAA ou TTTTTT). Dans ce type de séquence, la difficulté pour la technologie de séquençage Roche est de définir combien de fois la base répétée est présente. De plus, Roche a interrompu la commercialisation de cette technologie en 2016.

Illumina

La première machine utilisant la technologie Solexa (aujourd'hui Illumina) fut commercialisée en 2006. A l'origine, cette technologie ne produisait que des lectures très courtes (36 nucléotides). Aujourd'hui, les lectures sont en moyenne de 100 à 150 nucléotides avec un maximum de

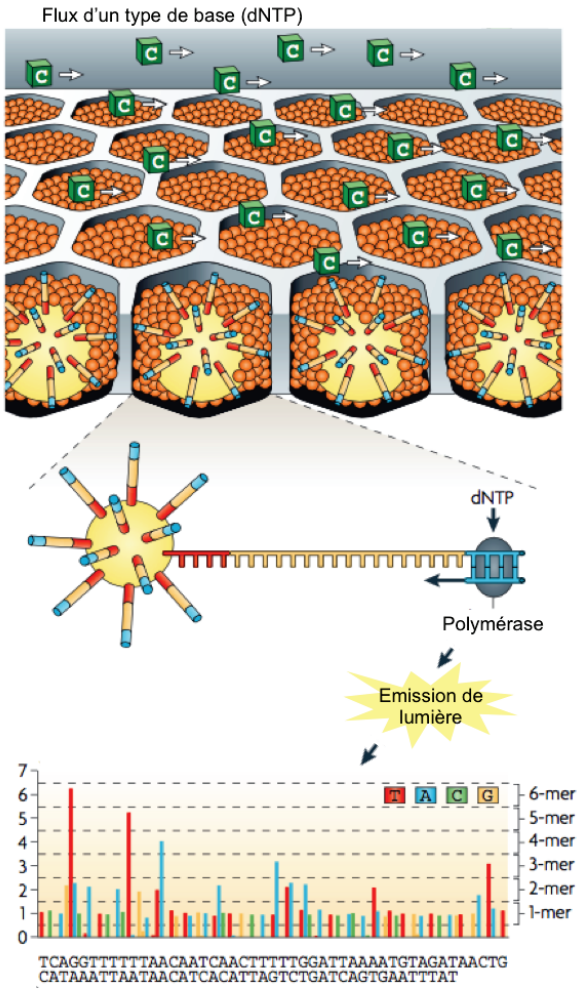


FIGURE 15 – **Le séquençage 454 ou pyroséquençage.** Chaque puits contient une bille sur laquelle un fragment a été amplifié. Le séquençage est réalisé par synthèse avec l'ajout successif des différents nucléotides modifiés. L'incorporation d'un nucléotide va permettre l'émission d'un signal lumineux. Ce signal est capté et interprété en termes de base nucléotidique. Plus le signal lumineux est important, plus le nombre de bases incorporées est grand. (Adaptée de [Metzker, 2010])

300 nucléotides pour le MiSeq. Les machines HiSeq X de chez Illumina sont celles qui actuellement permettent d'avoir la plus grande profondeur de séquençage (Figure 13). En effet, elles permettent d'obtenir jusqu'à 6 milliards de lectures en une seule passe de séquençage (run).

L'amplification clonale de la librairie d'ADN est réalisée par une PCR formant des ponts (Figure 14 B). Les fragments sous forme simple brin sont accrochés par une de leur extrémité à une surface recouverte de séquences adaptatrices. L'autre extrémité de chaque fragment est repliée et se lie à un adaptateur de la surface formant un pont. Cela permet d'initier la synthèse du brin complémentaire. De multiples cycles d'amplification suivi d'une dénaturation sont réalisés afin d'obtenir des clusters contenant environ 1000 copies de la molécule d'ADN simple brin d'origine. Par la suite, le séquençage Illumina est réalisé par synthèse avec l'utilisation des nucléotides

modifiés (Figure 16). Ces nucléotides sont des terminateurs réversibles de la synthèse et ils sont marqués par fluorescence (une couleur correspondant à une base). Après l'inclusion d'un nucléotide, une image du signal fluorescent est prise. Puis, le terminateur et le fluorochrome sont retirés et le cycle recommence pour pouvoir déterminer les bases suivantes. Le signal fluorescent est ensuite transformé informatiquement en séquence.

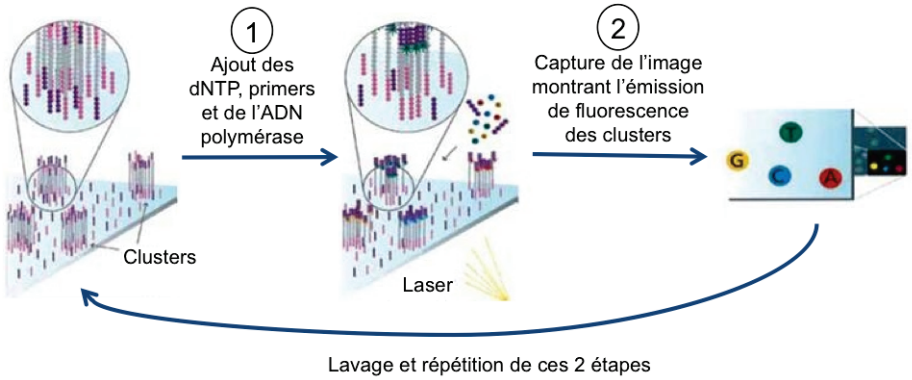


FIGURE 16 – **Le séquençage Illumina.** Une base est incorporée dans chaque cluster. Cette incorporation émet une fluorescence. Le signal fluorescent est enregistré sous forme d'image. Ces étapes sont répétées jusqu'à arriver à la longueur de lecture souhaitée.

Grâce à leur grande profondeur de séquençage, les séquençages Illumina et SOLiD furent pendant longtemps les méthodes les plus adaptées pour des analyses quantitative en transcriptomique (RNA-seq) ou en épigénétique (ChIP-seq). Illumina est actuellement le leader sur le marché du séquençage avec plus de 6 000 machines dans le monde.

SOLiD

Le premier séquenceur SOLiD fut commercialisé en 2007. Tout comme le séquençage Illumina, la technologie SOLiD permet d'obtenir uniquement des courtes lectures (maximum 75 nucléotides) avec une très grande profondeur (jusqu'à 3 milliards de lectures par run).

La méthode d'amplification est la même que celle des séquenceurs 454 : la PCR en émulsion (Figure 14 A). L'unique différence est qu'à la fin de l'amplification, les billes ne sont pas distribuées dans des puits mais elles sont fixées sur une plaque de verre.

Ensuite, le séquençage est réalisé en utilisant le principe de la ligation. Le processus consiste à ajouter grâce à une ligase des octamers (blocs de 8 nucléotides). Ces octamers (qui sont au nombre de 16) sont composés de 2 bases bien définies, 3 bases dégénérées et 3 bases universelles, comme représenté sur la figure 17. Chaque dinucléotide présent sur un octamer est associé à

une couleur de fluorescence (4 couleurs au total donc 4 dinucléotides correspondent à la même couleur). Après l'incorporation d'un octamer, un signal fluorescent est émis et enregistré. Les 3 dernière bases de cet octamer sont alors clivées et un nouvel octamer peut être incorporé. Ces étapes sont répétées jusqu'à arriver à la longueur de la lecture (au maximum 75 nucléotides). Puis, le brin complémentaire qui vient d'être synthétisé est éliminé et la synthèse recommence avec une nouvelle amorce décalée d'une base. Cette synthèse est effectuée avec 5 amorces différentes, permettant ainsi de déterminer la séquence nucléotidique du fragment.

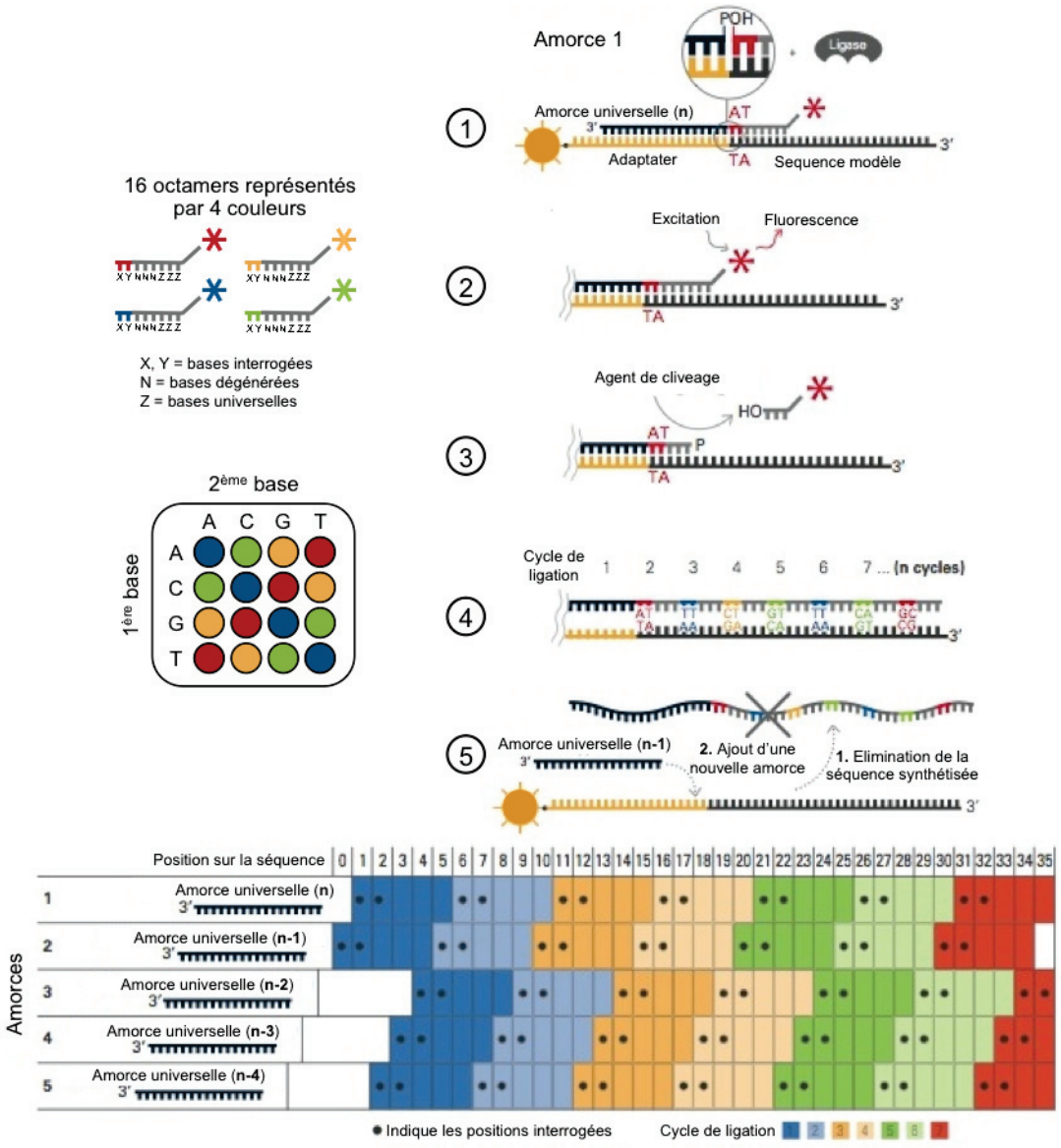


FIGURE 17 – **Le séquençage SOLiD.** La première étape est l'incorporation d'un octamer. Un signal fluorescent est enregistré puis les 3 dernières bases de l'octamer sont clivées. Ces 3 étapes sont répétées jusqu'à la taille de lecture souhaitée. Puis le brin qui vient d'être synthétisé est éliminé et une nouvelle amorce est associée au fragment modèle. La synthèse est réalisée avec 5 amorces différentes ce qui permet de lire 2 fois chaque base. (Adaptée de [Metzker, 2010])

Après la technologie Illumina, SOLiD produit la plus grande profondeur de séquençage. Par contre, la longueur des lectures qui est au maximum de 75 nucléotides reste une importante limite de cette technologie. Aujourd'hui, les séquenceurs SOLiD ne sont plus commercialisés.

Ion Torrent

La technologie Ion Torrent fut la dernière des technologies de seconde génération à être commercialisée, en 2010. Les lectures produites par cette méthode sont en moyenne de 200 nucléotides et au maximum de 400 nucléotides. Un run de séquençage peut générer jusqu'à 75 millions de lectures avec la dernière version de la machine S5 XL.

Ion Torrent utilise, comme les séquenceurs 454 et SOLiD, une amplification PCR par émulsion (Figure 14 A).

Le séquençage est réalisé par synthèse. Mais contrairement à toutes les autres technologies de séquençage décrites précédemment, Ion Torrent est la première utilisant un système de détection non optique. Comme pour le séquençage 454, les bases sont ajoutées séquentiellement avec un lavage avant chaque ajout d'une nouvelle base. Si une base est incorporée lors de la synthèse, un ion H^+ est libéré. Le pH du puits est mesuré pour déterminer si la base introduite à été incorporée. Plus la variation de pH mesurée est importante, plus le nombre de bases incorporées est grand. Cette mesure permet ainsi de déterminer la séquence de chaque fragment contenu dans un puits.

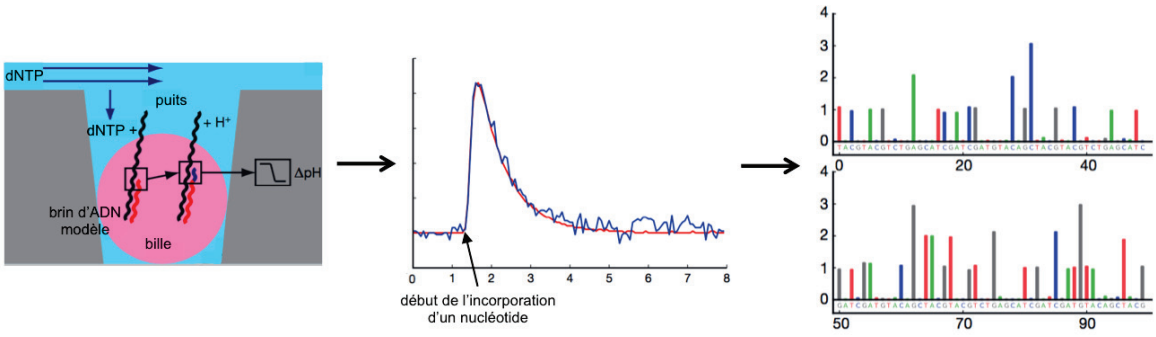


FIGURE 18 – **Le séquençage Ion Torrent.** Le premier schéma représente un puits dans lequel se trouve les fragments d'ADN fixés à une bille. Lors de l'incorporation d'un nucléotide, un ion H^+ est libéré. Cette libération d'ion modifie le pH du puits. Des capteurs mesurent ce changement de pH et le traduise en graphique, comme sur le schéma du milieu. Plus le changement de pH est être grand, plus un grand nombre de nucléotides ont été incorporés. Le dernier schéma représente les 100 premiers flux pour un puits (un flux étant l'introduction d'un nucléotide dans le puits). Chaque barre colorée indique le nombre de bases incorporées pendant le flux du nucléotide en question. (Adaptée de [Rothberg et al., 2011])

Tout comme les séquenceurs 454, cette technologie a un fort taux d'erreur dans les répétitions d'homopolymères. Le principal avantage de cette technologie est la vitesse des runs. En effet, un run typique ne prend que quelques heures, comparé à un ou plusieurs jours pour la technologie Illumina.

3 Les technologies de séquençage de troisième génération

Aujourd'hui, il existe une dernière génération de technologies de séquençage produisant de très longues lectures (Figure 13) : PacBio (Pacific Biosciences) et MinION de Oxford Nanopore. Ces technologies, dites de troisième génération, permettent de séquencer des molécules complètes sans amplification préalable. Mais ces technologies de séquençage manquent encore de profondeur et donc restent actuellement très chères au niveau du coût par base séquencée. De plus, elles ont un taux d'erreur beaucoup plus élevé que les technologies de seconde génération.

PacBio

La technologie de PacBio est commercialisée depuis 2010. Cette technologie est la première des technologies de troisième génération. Elle ne nécessite pas d'amplification clonale de l'ADN avant le séquençage. La longueur des molécules fournies au séquenceur définit la longueur maximale des lectures. Cette propriété donne ainsi la possibilité d'avoir de très longues lectures.
Cette technologie réalise du séquençage en temps réel. Il n'y a en effet pas d'arrêt du processus de synthèse de l'ADN. Une seule molécule d'ADN polymérase est fixée au fond de chaque puits (Figure 19 A) permettant de contrôler la synthèse du brin d'ADN complémentaire. C'est le brin d'ADN modèle qui va se déplacer base par base sur cette ADN polymérase. Des nucléotides marqués par fluorescence sont intégrés dans l'ADN. Des capteurs intégrés dans les puits vont recevoir le signal fluorescent émis par l'incorporation du nucléotide. Chaque nucléotide étant marqué avec une couleur différente, le signal peut être traduit en base (A, T, G, C) (Figure 19 B).

L'incorporation des nucléotides étant très rapide (séparée par quelques millisecondes), un run de séquençage de PacBio s'effectue en quelques heures seulement. Mais, cette technologie reste encore limitée par son fort taux d'erreur (entre 11% et 14%) et sa faible profondeur de séquençage. En effet, un run de séquençage produit environ 55 000 lectures et même si ces lectures sont très longues, des séquences présentes en très faible quantité dans l'échantillon de départ ont peu de

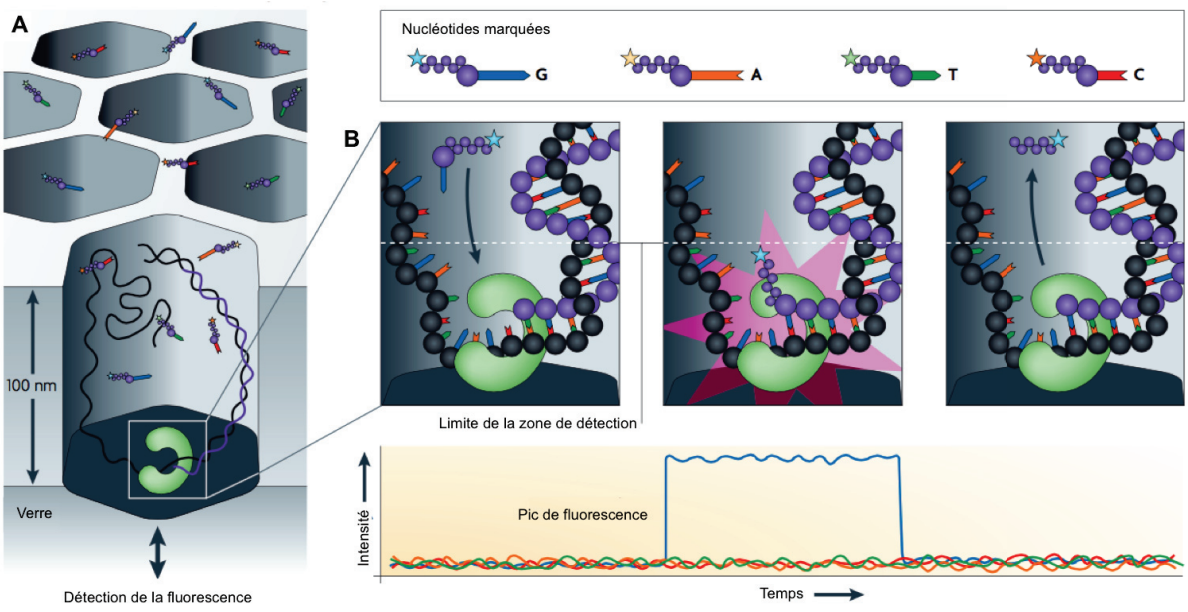


FIGURE 19 – **Le séquençage PacBio.** Le séquençage est réalisé en temps réel à l'aide d'une ADN polymérase fixée au fond de chaque puits et de nucléotides marqués par fluorescence. Lors de l'incorporation d'un nucléotide pendant la synthèse du brin complémentaire, un signal fluorescent est émis. Celui-ci est capté et interprété grâce au marquage spécifique de chaque base. (Adaptée de [Metzker, 2010])

chance d'être séquencées. La limitation de cette technologie se trouve donc dans le nombre de lectures et non dans le nombre de bases séquencées.

Oxford Nanopore

La dernière technologie de séquençage présentée ici est la technologie développée par Oxford Nanopore. Le MinION, un séquenceur à peine plus gros qu'une clé usb, est un séquenceur utilisant un pore biologique pour réaliser le séquençage (Figure 20 A). Ce type de séquenceur génère environ 10 mégabases par run de 16h et les lectures font en moyenne environ 6 000 bases. Cette machine est actuellement la moins chère du marché .

Les pores biologiques utilisés dans cette technologie sont retenus dans une double membrane lipidique. Sur la molécule d'ADN, deux adaptateurs sont fixés. Le premier est lié à une protéine motrice qui va faire passer la molécule d'ADN dans le pore. Le deuxième crée la boucle reliant les 2 brins complémentaires du fragment d'ADN. Il est également lié à une protéine motrice. Ces deux brins d'ADN complémentaires sont séquencés, permettant ainsi d'augmenter la précision des résultats. Le séquençage est réalisé en mesurant les changements de courant induit par le passage des nucléotides dans le pore (Figure 20 B), chaque nucléotide ayant une signature spécifique.

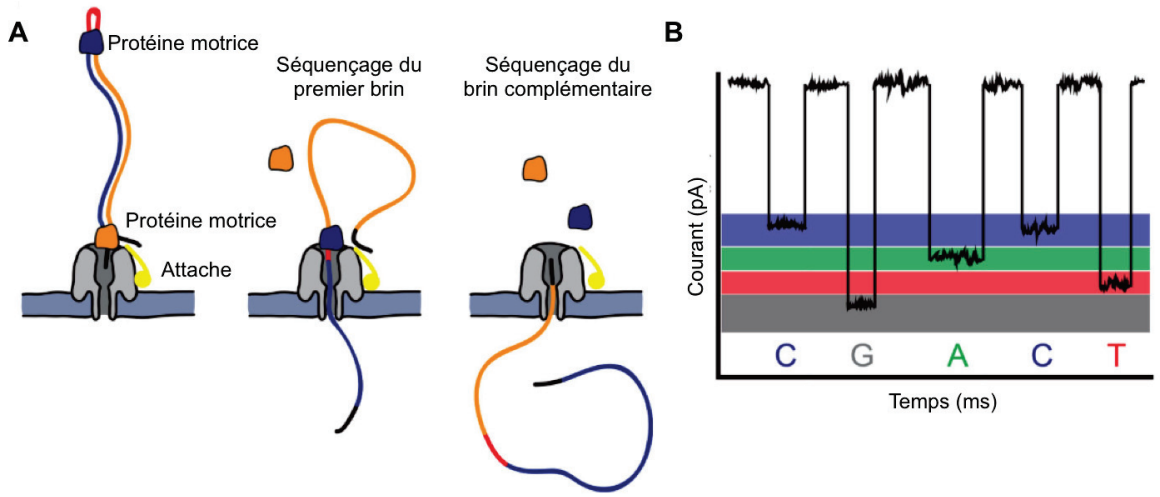


FIGURE 20 – **Le séquençage Oxford Nanopore.** Le séquençage est réalisé, comme pour PacBio, en temps réel. **A.** Les 2 brins complémentaires d'un fragment d'ADN reliés par un adaptateur passent dans un pore biologique, dirigés par des protéines motrices. **B.** Le courant produit par le passage de chaque nucléotide dans le pore est enregistré et interprété pour déterminer la séquence du fragment. (Adaptée de [Reuter et al., 2015])

Comme pour la technologie PacBio, les limites du séquenceur Oxford Nanopore restent la faible profondeur de séquençage et le fort taux d'erreur (entre 10 et 15%). En effet, le MinION génère environ 4.4 millions de bases en un run de séquençage en mode rapide contre 1.5 milliards de bases pour le HiSeq de chez Illumina.

B Les différentes applications

Les trois grands domaines d'application du séquençage haut-débit sont la génomique, la transcriptomique et l'épigénétique (Figure 21).

Depuis l'apparition des technologies NGS, de nombreux types de séquençage ont été développés pour répondre à des questions biologiques particulières. La première application est le séquencage de génomes complets. Ce type de séquençage sert à déterminer *de novo* la séquence du génome pour des organismes dont on ne connait pas encore le génome. Avec le reséquençage de génome, on peut déterminer les SNV (variations d'un seul nucléotide) dans une population et les réarrangements chromosomiques et les variations de la structure d'un génome. Le séquençage de l'exome (séquençage des fragments correspondant aux exons uniquement) permet aussi d'adresser les questions de détermination de SNV. De plus, cette méthode permet d'obtenir une couverture des exons beaucoup plus élevée avec une profondeur de séquençage moins importante. La dernière application du séquençage de génome est la métagénomique. Dans ce type d'étude,

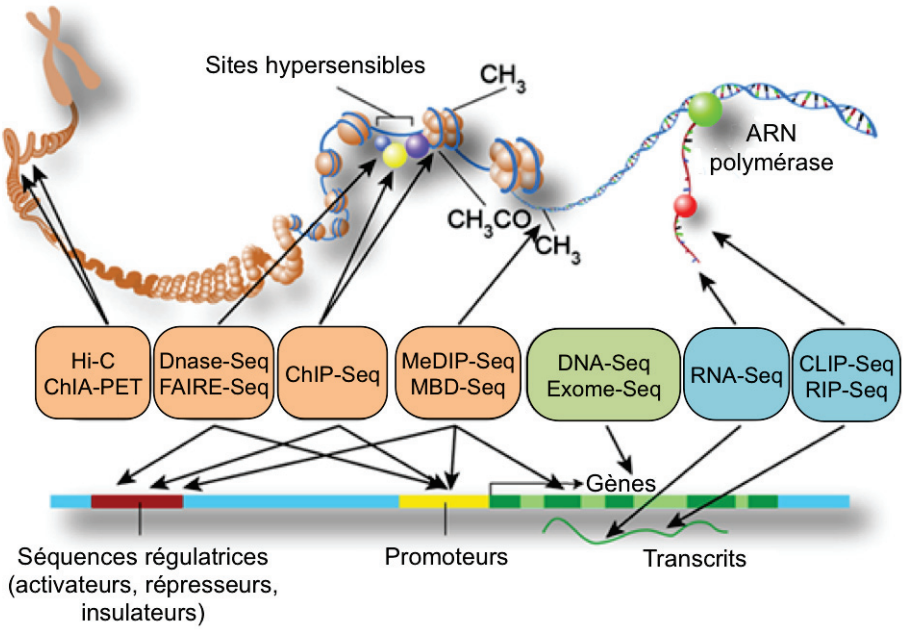


FIGURE 21 – **Les applications des NGS.** Le séquençage haut-débit s'applique aussi bien à la génomique (cadre vert), qu'à la transcriptomique (cadres bleus) et à l'épigénétique (cadres oranges). (Adaptée d'ENCODE)

l'ADN de tous les organismes vivant dans un milieu est extrait et séquencé. La métagénomique est notamment beaucoup utilisé pour étudier le microbiome humain (les micro-organismes présents dans le corps humain).

De nombreuses méthodes de séquençage ont pour but d'identifier les régions régulatrices du génome (Tableau 1). Le ChIP-Seq est la première de ces méthodes [Johnson et al., 2007]. Elle permet d'identifier les régions du génome sur lesquelles des protéines (comme des facteurs de transcription, d'épissage ou des histones avec des modifications bien spécifiques) se lient. Une méthode plus générale pour déterminer les potentielles régions régulatrices est de cartographier les régions "ouvertes" de la chromatine en utilisant une enzyme de digestion, la DNase I. Les fragments récupérés peuvent être ensuite séquencés. Cette méthode est appelée DNase-Seq [Song and Crawford, 2010]. Mais le DNase-Seq est peu à peu remplacé par la technique d'ATAC-Seq [Buenrostro et al., 2013]. En effet, le protocole d'ATAC-Seq est plus simple et nécessite moins de matériel de départ. Ces méthodes permettent aussi d'identifier les régions sur lesquelles les facteurs de transcription se lient [Tsompana and Buck, 2014]. Une autre méthode, appelée FAIRE-Seq, permet d'étudier les régions du génome associées avec une activité régulatrice.

Le séquençage haut-débit est également beaucoup utilisé en transcriptomique. Le séquençage des

Methode	Description	Référence
ChIP-Seq (Chromatin immunoprecipitation sequencing)	Identification les sites de fixations de protéines cibles sur l'ADN.	[Johnson et al., 2007]
DNase-Seq (DNase I hypersensitive sites sequencing),	Identification les régions "ouvertes" de la chromatine.	[Song and Crawford, 2010],
ATAC-Seq (Assay for transposon accessible chromatin sequencing) et		[Buenrostro et al., 2013]
FAIRE-Seq (Formaldehyde-assisted isolation of regulatory elements sequencing)		[Giresi and Lieb, 2009]
ChIA-PET (Chromatin interaction analysis by paired-end tag sequencing) et	Etude de la structure 3D du génome.	[Fullwood and Ruan, 2009]
Hi-C (High chromosome contact map)		[Lieberman-Aiden et al., 2009]
MeDIP-Seq (Methylated DNA Immunoprecipitation sequencing),	Etude de la méthylation de l'ADN.	[Jacinto et al., 2008]
MRE-Seq (Methylation-sensitive Restriction Enzyme sequencing) et		[Maunakea et al., 2010]
MBD-Seq (Methylated DNA Binding Domain sequencing)		[Serre et al., 2010]

TABLE 1 – Liste non exhaustive de méthodes de séquençage haut-débit appliquées à l'épigénétique avec une courte description.

ARN, appelé RNA-Seq, est devenu la méthode classique pour les analyses transcriptomiques. Une explication plus approfondie du RNA-seq sera donnée dans la section III.B. Des combinaisons de techniques biologiques et biochimiques avec du séquençage permettent d'analyser différents aspects de la transcription (Tableau 2). Les techniques de GRO-Seq, PRO-Seq ou NET-Seq permettent de séquencer les ARNs en cours de transcription. Le Ribo-Seq et le TRAP-Seq permettent d'analyser les ARNs en cours de traduction. Le RIP-Seq, le CLIP-Seq (appelé aussi HITS-CLIP), le PAR-CLIP et l'iCLIP permettent d'étudier les sites de fixations de protéines cibles sur l'ARN (équivalent du ChIP-Seq pour l'ARN). La technique de ChIRP-Seq permet d'étudier les interactions ADN-ARN. En effet, elle permet de déterminer les régions de la chromatine qui interagissent avec des ARNs spécifiques. La méthode de PARE-Seq permet de déterminer les sites de clivage des micro ARNs et d'étudier la dégradation des ARNs.

Le séquençage NGS a également été utilisé pour étudier d'autres aspects de l'épigénétique comme la méthylation de l'ADN (Tableau 1). Différentes méthodes ont été développées pour réaliser ces analyses, comme par exemples le MBD-Seq, le MeDIP-Seq ou le MRE-Seq.

Enfin, le séquençage haut-débit permet également d'adresser la question de la conformation du génome. En effet, les méthodes de ChIA-PET et de Hi-C permettent d'étudier la conformation de la chromatine à l'échelle de tout le génome.

Cette liste n'est pas exhaustive, car il existe beaucoup d'autres méthodes qui ont été développées pour des questions biologiques bien précises.

Méthode	Description	Référence
mRNA-seq	Identification les ARN messagers.	[Mortazavi et al., 2008]
miRNA-seq	Identification les micro ARN.	[Ruby et al., 2006]
GRO-Seq (Global Run-On Sequencing),	Sélection et séquençage uniquement les ARNs en cours de transcription par l'ARN polymérase II.	[Core et al., 2008]
PRO-Seq (Precision Run-On Sequencing) et		[Kwak et al., 2013]
NET-Seq (Native elongation transcript sequencing)		[Churchman and Weissman, 2011]
Ribo-Seq (Ribosome profile sequencing)	Identification les ARNs messagers en cours de traduction.	[Ingolia et al., 2009]
et TRAP-Seq (Targeted purification of polysomal mRNA sequencing)		[Reynoso et al., 2015]
RIP-Seq (RNA immunoprecipition sequencing),	Détermination des régions d'ARN liées à une protéine d'intérêt.	[Cloonan et al., 2008]
CLIP-Seq (Cross-linking and immunoprecipitation sequencing),		[Chi et al., 2009]
PAR-CLIP (Photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation) et		[Hafner et al., 2010]
iCLIP (individual-nucleotide resolution CLIP)		[Huppertz et al., 2014]
ChIRP-Seq (Chromatine isolation by RNA purification)	Identification des régions du génome qui interagissent avec l'ARN.	[Chu et al., 2011]
PARE-Seq (Parallel analysis RNA ends sequencing)	Etude des sites de clivage des micro-ARNs ainsi que de la dégradation des ARNs.	[German et al., 2009]

TABLE 2 – Liste non exhaustive de méthodes de séquençage haut-débit appliquées à la transcriptomique avec une courte description. (Adaptée de [Anamika et al., 2016])

C L'analyse des données haut débit

À la sortie du séquenceur, l'utilisateur obtient un fichier au format fasta ou fastq (Figure 22). Ces fichiers sont des fichiers texte contenant pour chaque lecture un identifiant, la séquence nucléotidique et la qualité de séquençage (pour le fichier fastq). Ces fichiers contenant plusieurs millions de lectures, sont très volumineux. Le traitement de ces millions de lectures est donc impossible à faire manuellement. L'informatique s'est donc imposée comme une nécessité pour les biologistes, aussi bien pour le stockage des données que pour leur traitement.

Afin de gérer cette masse de données générée par les séquenceurs, les bioinformaticiens et les biostatisticiens ont du développer de nouveaux outils, rapides et fiables. Comme nous l'avons vu dans la section précédente, les domaines d'applications du séquençage NGS sont très variés et chaque application nécessite une analyse dédiée. Les bioinformaticiens doivent donc constamment adapter leurs méthodes aux nouvelles applications du séquençage ou en développer de nouvelles.

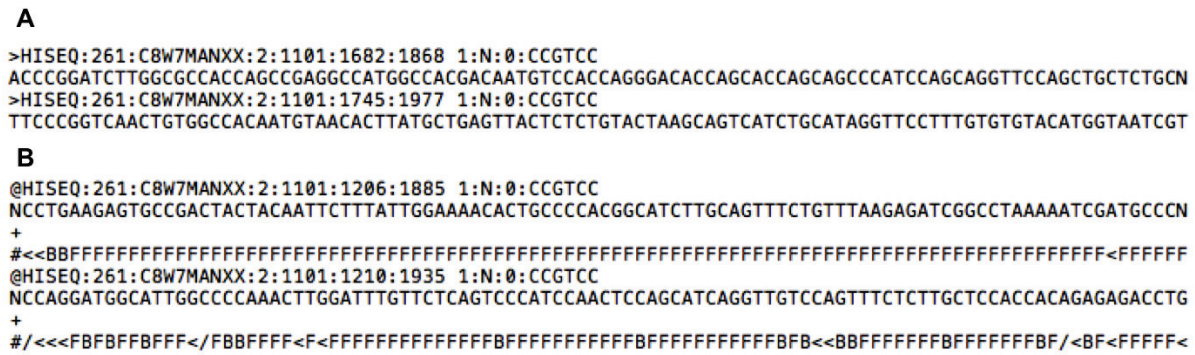


FIGURE 22 – **Les formats fasta et fastq. A.** Le format fasta permet de stocker des séquences biologiques, comme les lectures provenant de séquenceur. Chaque séquence est représentée par une première ligne débutant par un chevron ">", suivi de l'identifiant de la séquence. Les lignes suivantes contiennent la séquence nucléotidique, jusqu'à ce qu'une ligne commence par un chevron. **B.** Le format fastq est similaire au format fasta, mais il permet non seulement de stocker des séquences biologiques mais aussi leur score de qualité. Ce format représente une séquence avec 4 lignes. La première ligne débute par le symbole "@", suivi de l'identifiant de la séquence. La deuxième ligne représente la séquence nucléotidique. La troisième ligne commence par le symbole "+", parfois suivi de l'identifiant de la séquence (comme sur la première ligne). Enfin, la quatrième ligne contient le score de qualité associé à chacune des bases de la séquence nucléotidique. Le nombre de caractère présent sur cette ligne doit être égal au nombre de caractère présent dans la séquence nucléotidique de la deuxième ligne.

De plus, la puissance de calcul nécessaire pour traiter ces données a dépassé les capacités de nos ordinateurs de bureau. L'utilisation de cluster de calcul est donc devenue incontournable. Un cluster de calcul est un regroupement d'ordinateurs permettant de dépasser les limitations d'une machine en terme de capacité de calcul. Ce changement d'infrastructure a nécessité une modification des outils, mais l'inverse est vrai également. En effet, les particularités de la bioinformatique, en terme de traitement de données, ont dû engendrer une adaptation au sein des clusters de calcul, notamment avec une demande plus importante en mémoire vive.
Le stockage des données pose un véritable problème, car de plus en plus de données sont générées, augmentant constamment la demande en espace de stockage.

D Apports des données large échelle

Les avancées des technologies NGS ont permis aux chercheurs d'étudier les systèmes biologiques de façon beaucoup plus précise et rapide. Avec la démocratisation de ces technologies, de plus en plus de projets de grande envergure ont été entrepris. Notamment le projet 1 000 Génomes [1000 Genomes Project Consortium et al., 2012, Sudmant et al., 2015] et le projet de séquençage d'exome "Exome Sequencing Project" [Tennessen et al., 2012, Fu et al., 2013] ont produit une très grande quantité de données et modifié la façon de faire de la génomique. Ces

études ont permis de décrire des millions de variants dans plusieurs milliers d'individus provenant de différentes populations.

Deux autres projets de plus grande envergure sont annoncés : le projet 100 000 Génomes en Angleterre et le projet GenomeAsia 100K qui prévoit également de séquencer 100 000 individus. Ces projets doivent permettre de mieux décrire certaines populations ou ethnies qui ne sont actuellement pas bien étudiées. En effet, le projet GenomeAsia 100K a pour but de séquencer des individus provenant de 12 pays du sud de l'Asie et d'au moins 7 pays du nord et de l'est de l'Asie. Un des objectifs est de produire un génome de référence pour les populations asiatiques et d'identifier les allèles rares et fréquents associés à cette population. Ces projets ont aussi un objectif médical.

Les NGS sont devenus des technologies clés en recherche fondamentale et sont également en train de devenir des outils incontournables dans la recherche translationnelle et dans le diagnostique. En effet, le séquençage haut-débit a permis la création de bases de données regroupant un très grand nombre de jeux de données sur certaines maladies. Pour le cancer, on peut donner l'exemple de TCGA (The Cancer Genome Atlas) ou d'ICGC (International Cancer Genome Consortium). TCGA est une très grande base de données contenant 2,5 petabytes de données provenant de tissus tumoraux et de tissus sains contrôle de plus de 11 000 patients. Ces données permettent d'étudier 33 types de cancer grâce à la disponibilité des données pour la communauté scientifique. Les données post-traitement bioinformatique sont librement accessibles sur leur site donnant ainsi accès aux plus petits laboratoires à une masse de données qu'ils ne pourraient pas générer eux même. De plus, le projet 100 000 Génomes dont nous avons parlé précédemment doit séquencer le génome d'individus atteints de maladie rares ou de cancer. Ce type de projet a pour objectif de mener vers une médecine personnalisée. En effet, en caractérisant la tumeur de chaque patient, on peut choisir le traitement le plus adapté à chacun.

Enfin, le séquençage peut générer des données en génomique, en transcriptomique et en épigénétique. L'intégration de ces différents niveaux d'information donne une image encore plus précise du fonctionnement de la cellule et permet d'aller encore plus loin dans la compréhension des mécanismes complexes de régulation de l'expression des gènes ou de l'épissage alternatif.

III Méthodes d'analyse de l'épissage

A Historique de l'analyse de l'épissage : du cas par cas au large échelle

Pendant longtemps, les études sur l'épissage alternatif étaient réalisées par RT-PCR au cas par cas. Ce travail était long et fastidieux.

L'informatique a changé les perspectives en matière d'expérimentation biologique. Les avancées en bioinformatique ont fait évoluer les techniques d'analyse grâce à l'automatisation, accélérant ainsi les découvertes des évènements d'épissage alternatif.

Parmi ces technologies permettant de faire du large échelle, on trouve les technologies de PCR à haut débit, de puce à ADN et de séquençage haut débit. Chacune de ces technologies a des avantages et des inconvénients pour les analyses à large échelle de l'épissage alternatif. Mais toutes n'ont été concevables que grâce à l'introduction de l'informatique au sein de la biologie.

1 La RT-PCR

La RT-PCR (reverse transcriptase polymerase chain reaction) permet d'amplifier de l'ARN en ADN complémentaire (ADNc). L'ARN est d'abord rétrotranscrit en ADNc et ce dernier va être amplifié comme lors d'une PCR classique à partir d'amorces conçues au préalable.

Cette technique permet d'analyser l'épissage alternatif grâce à une conception spécifique des amorces dans les exons flanquants l'événement d'épissage (Figure 23). Cette méthode, qui est toujours utilisée aujourd'hui pour valider les prédictions faites à large échelle, présente certains désavantages. Tout d'abord, l'analyse d'événements d'épissage avec ce type de technique doit être faite au cas par cas. En effet, des amorces doivent être conçues pour chaque événement que l'on souhaite tester. De plus, la conception de ces amorces nécessite une connaissance préalable du gène et de son épissage. Qui plus est, ces informations sont également nécessaires pour vérifier que les produits de la PCR (les amplicons) sont de la taille attendue.

2 La PCR haut débit

La PCR à haut débit a été développée par des chercheurs de l'Université de Sherbrooke au Québec sous le nom de LISA (Layered and Integrated System for Splicing Annotation). Ils l'ont utilisée pour identifier les marqueurs d'épissage alternatif du cancer du sein [Venables et al., 2008] et ceux du cancer des ovaires [Klinck et al., 2008].

Le principe de cette PCR à haut débit est de réaliser plusieurs RT-PCR pour analyser un évé-

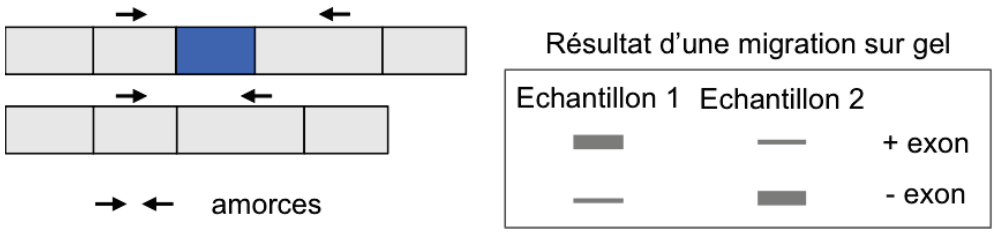


FIGURE 23 – **La technique de RT-PCR.** Les amorces doivent être conçues pour avoir une séquence complémentaire de séquence des exons flanquants l'événement d'épissage (ici l'exon exclu). Après migration de l'ADN obtenu sur gel d'agarose, on obtient autant de bandes qu'il y a de variants. Dans le cas schématisé, il y a 2 variants. Les bandes les plus hautes correspondant aux amplicons les plus grands, la bande du haut correspond à l'inclusion de l'exon et celle du bas à son exclusion. (Adaptée de [Wang and Cooper, 2007])

nement d'épissage et de nombreux événements peuvent être analysés en même temps. Cette technique automatise la méthode d'analyse de gènes candidats grâce à l'informatique et à la robotique.

La plateforme LISA génère les cartes des transcripts pour chaque gène sélectionné à partir de bases de données publiques. Elle détermine les expériences de PCR à faire afin que chaque jonction d'exon supposée et chaque évènement d'épissage possible soient couverts par au moins 2 réactions de PCR. Ces RT-PCR sont ensuite réalisées et analysées.

La plateforme LISA est capable de réaliser et d'analyser 3 000 réactions par jour. L'importante quantité d'informations obtenue en peu de temps est une avancée majeure par rapport aux analyses gène à gène. De plus, cette technique ne nécessite pas de modèles mathématiques complexes pour interpréter les résultats et comporte moins de biais que certaines méthodes (comme la technologie de RNA-seq Illumina). Mais elle nécessite une connaissance préalable sur les gènes que l'on veut étudier et donc, il est impossible d'identifier de nouvelles jonctions.

Grâce à cet outil, l'équipe de l'Université de Sherbrooke a pu déterminer des évènements d'épissage alternatif associés au cancer du sein et d'autres associés au cancer des ovaires. Les résultats ont été encourageant, car dans les deux études les évènements d'épissage alternatif trouvés ont permis de classifier les échantillons comme cancéreux ou normaux avec un faible taux d'erreur [Venables et al., 2008, Klinck et al., 2008].

3 Les puces à ADN

Une puce à ADN est une petite plaque quadrillée par des millions de séquences d'oligonucléotides synthétiques (sondes) correspondant à un génome d'intérêt. On connait à l'avance à quoi correspondent les sondes et leur emplacement sur la grille.

Le principe d'une puce à ADN est de faire s'hybrider des brins d'ADN complémentaires (ADNc) obtenus à partir d'ARNm par transcription inverse, sur les sondes de la puce. Les brins d'ADNc sont marqués par fluorescence avant l'étape d'hybridation. On peut donc détecter les endroits où les séquences ce sont hybridées grâce à une caméra qui enregistre l'intensité de fluorescence. L'analyse bioinformatique des intensités permet ensuite d'identifier les ARNm présents dans l'échantillon et de comparer plusieurs échantillons entre eux (Figure 24). Cette technique permet de se libérer de certaines limitations des études réalisées à partir d'ESTs, comme le fort biais vers l'extrémité 3' du gène. Les ESTs ("Expressed Sequence Tag") sont de courtes séquences d'ADNc qui ont longtemps servies à identifier les gènes et les transcrits exprimés dans les cellules. Les puces ont ainsi permis de trouver de nombreux nouveaux évènements d'épissage [Lee and Wang, 2005].

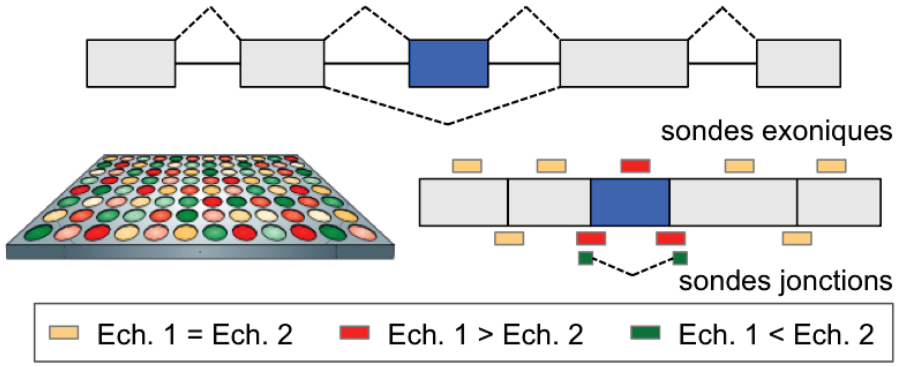


FIGURE 24 – **La puce à ADN.** La puce à ADN permet de déterminer les ARNs présents dans un échantillon grâce à des sondes qui ont été conçues dans les exons et/ou sur les jonctions d'exons (suivant la génération de la puce). Grâce à l'analyse bioinformatique, les échantillons peuvent être comparés entre eux. Classiquement, une sonde qui a montré une intensité plus importante dans un des deux échantillons comparés est représentée en rouge ou en vert. Ech. = échantillon. (Adaptée de [Wang and Cooper, 2007])

Il existe trois générations de puces (chez Affymetrix) : la puce classique, la puce exon et la puce jonction.

Avec la puce classique, appelée aussi puce 3', l'expression du gène est déduite à partir de l'expression de son extrémité 3'. Elle détecte des différences de niveau d'expression entre différents tissus ou différentes conditions. Mais elle n'est pas du tout adaptée à l'analyse de l'épissage alternatif car elle ne permet pas de différencier les transcrits.

La puce exon permet d'étudier les différents exons séparément grâce à des sondes spécifiques à chacun d'eux. Ces puces permettent d'analyser l'épissage alternatif mais elles restent limitées par le faible nombre de sondes ciblant chaque exon. En effet, un exon étant en moyenne représenté par quatre sondes, les résultats dépendent très fortement de la qualité de l'hybridation et du

marquage fluorescent.

La puce jonction est la dernière génération de puce Affymetrix. En plus des sondes exons, elle possède des sondes spécifiques aux jonctions d'exons. De plus, chaque exon est représenté de manière beaucoup plus importante (10 sondes par exon). L'analyse de l'épissage alternatif avec ce type de puce est donc plus fiable et plus précis.

[Dutertre et al., 2010] ont utilisé des puces jonctions pour étudier des cellules mammaires tumorales avec une plus ou moins grande capacité à se disséminer. Ils ont trouvé des évènements d'épissage alternatif liés à des mauvais pronostics dans une cohorte de patientes atteintes de cancer du sein. Cette technologie a été utilisée dans de très nombreuses études, notamment dans certaines concernant le cancer, pour trouver de nouveaux marqueurs de pronostiques et développer de nouveaux outils thérapeutiques. Mais elle est actuellement largement remplacée par les techniques de séquençage haut débit.

B Le séquençage du transcriptome : RNA-Seq

Comme nous l'avons vu dans la section II.B, il existe de très nombreuses applications au séquençage haut débit. Le RNA-Seq est l'une d'entre elles.

Le RNA-Seq consiste à extraire et séquencer le transcriptome de cellules. Il existe de nombreux types d'expériences de RNA-Seq. Le choix du type de librairie à réaliser dépend de la question biologique. Le plus courant est de séquencer uniquement les longs ARNs (supérieurs à 200 nucléotides) en réalisant une sélection de taille. Lorsque uniquement les petits ARNs (inférieurs à 200 nucléotides) sont séquencés, on parle de small RNA-Seq. Une seconde sélection est appliquée sur les longs ARN avant séquençage. En effet, il faut éliminer les ARNs ribosomiques (ARNr), qui représentent la majorité (>90%) de l'ARN total d'une cellule, pour éviter de "gaspiller" de la profondeur de séquençage. Pour réaliser cette sélection, différents protocoles sont disponibles. Le protocole de Ribo-zero consiste à éliminer uniquement l'ARNr de l'ARN total. Un autre protocole, appelé polyA+, permet de ne sélectionner que les ARNs possédant une queue polyA. Enfin, il est également possible de ne sélectionner que les ARNs ne possédant pas de queue polyA, en réalisant une librairie polyA-.

D'autres choix doivent être faits lors de la construction de la librairie. On peut réaliser des librairies *single-end* ou *paired-end*. Dans les librairies *single-end* uniquement une des deux extrémités du fragment est séquencée. Alors que les deux extrémités de ce fragment seront séquencés avec une librairie *paired-end*. Enfin, la librairie peut être brin spécifique. C'est-à-dire que l'information

sur le brin d'origine des gènes est conservée. Il existe plusieurs types de librairie brin spécifique : soit les lectures produites s'aligneront sur le génome dans le même sens que le gène, soit dans le sens opposé au sens du gène.

Le principal avantage du séquençage RNA-Seq par rapport aux puces à ADN est qu'aucune connaissance préalable sur le transcriptome n'est nécessaire. Contrairement aux puces, le RNA-Seq permet donc de découvrir de nouveaux exons, transcrits ou événements d'épissage. Cette technologie a donc permis d'aller beaucoup plus loin dans l'analyse du transcriptome.

Le séquençage d'ARN est un outil très puissant car il permet d'adresser de nombreuses questions. A partir de données RNA-Seq, il est possible de déterminer l'expression et l'épissage des gènes mais aussi les mutations ponctuelles dans la région codante des gènes ou encore les fusions de transcrits. Mais pour chacune de ces analyses, le design recommandé n'est pas le même. Notamment pour l'analyse de l'épissage alternatif, il est important de ne pas avoir des lectures trop courtes. En effet, plus les lectures vont être longues, plus il est probable qu'elles chevauchent une jonction d'exons donnant ainsi une information importante pour l'analyse de l'épissage. Il est donc recommandé de séquencer des lectures d'au moins 100 nucléotides de longueur. Il est également recommandé d'utiliser une librairie brin spécifique. En effet, lorsque des gènes se chevauchent sur le génome et sont sur les brins opposés de l'ADN, il est difficile de déterminer quelle lecture provient de quel gène. En réalisant une librairie brin spécifique, on s'affranchit de ce problème et on évite ainsi de trouver de faux événements d'épissage qui pourraient en découler. Enfin, pour réaliser une analyse d'épissage alternatif chez l'homme, il est recommandé de séquencer au minimum 50 million de lectures *paired-end* pour avoir suffisamment de puissance à la fois pour la détection des événements d'épissage et pour l'analyse différentielle [Liu et al., 2013].

C L'analyse de l'épissage dans les données RNA-seq

On se concentrera dans cette section sur les méthodes bioinformatiques d'analyse de données RNA-Seq permettant d'étudier l'épissage alternatif. Il existe de nombreux autres outils pour analyser ces données sous différents angles : analyse de l'expression des gènes, détection de SNPs (polymorphisme d'un nucléotide) ou de fusion de gènes. Il faut aussi noter que certains des outils présentés dans cette section permettent d'analyser d'autres aspects des données RNA-Seq, comme l'expression des gènes.

1 Pipeline d'analyse de l'épissage alternatif

Ces outils peuvent être catégorisés selon plusieurs critères : méthodes basées sur l'alignement ou sur l'assemblage, méthodes locales ou globales.

La première distinction que l'on peut faire est entre les méthodes basées sur l'alignement et celles basées sur l'assemblage. Les méthodes basées sur l'alignement sont les plus couramment utilisées. Les lectures sont assignées aux positions du génome de référence desquelles elles peuvent provenir. Ces méthodes nécessitent un génome de référence et également un transcriptome de référence pour obtenir un meilleur résultat lors de l'étape d'alignement. Leur principal point limitant est la gestion des lectures s'alignant à plusieurs loci sur le génome. La solution la plus utilisée est d'exclure ces lectures de l'analyse. Les méthodes basées sur l'assemblage peuvent tout aussi bien être utilisées pour des organismes modèles (possédant un génome de référence) ou pour des organismes ne possédant pas de génome de référence. Ces méthodes restent peu utilisées lorsqu'un génome de référence est disponible, essentiellement pour des raisons historiques. En effet, ces méthodes demandaient beaucoup de temps et de ressources de calcul. Dans le cas où un génome de référence est disponible, l'alignement est réalisé avec les séquences assemblés. Ces séquences étant plus longues que les lectures, il y a moins de séquences pouvant s'aligner à plusieurs loci sur le génome.

La deuxième séparation pouvant être faite dans les méthodes d'analyse est entre les méthodes locales et les méthodes globales. Les méthodes locales sont focalisées sur les exons, les morceaux d'exons ou les événements d'épissage. Elles ne cherchent pas à reconstruire les transcrits complets mais seulement à extraire les événements d'épissage. Ces méthodes peuvent être basées sur l'alignement des lectures comme MISO ou DEXseq ou basées sur l'assemblage de ces lectures comme KisSplice. Au contraire, certaines méthodes ont pour objectif de reconstruire les transcrits complets et ce sont ces transcrits qui sont alors comparés lors de l'analyse différentielle. Comme pour les méthodes locales, il existe des méthodes globales basées sur l'alignement comme Cufflinks et d'autres basées sur l'assemblage comme Trinity.

Les principales étapes d'une analyse de l'épissage alternatif dans les données RNA-Seq sont schématisées dans la figure 25.

2 Contrôle qualité

Quelques soit la type d'analyse souhaité, la première étape à réaliser sur les données RNA-Seq est le contrôle qualité. Différents outils permettent de vérifier que les données sont de bonne

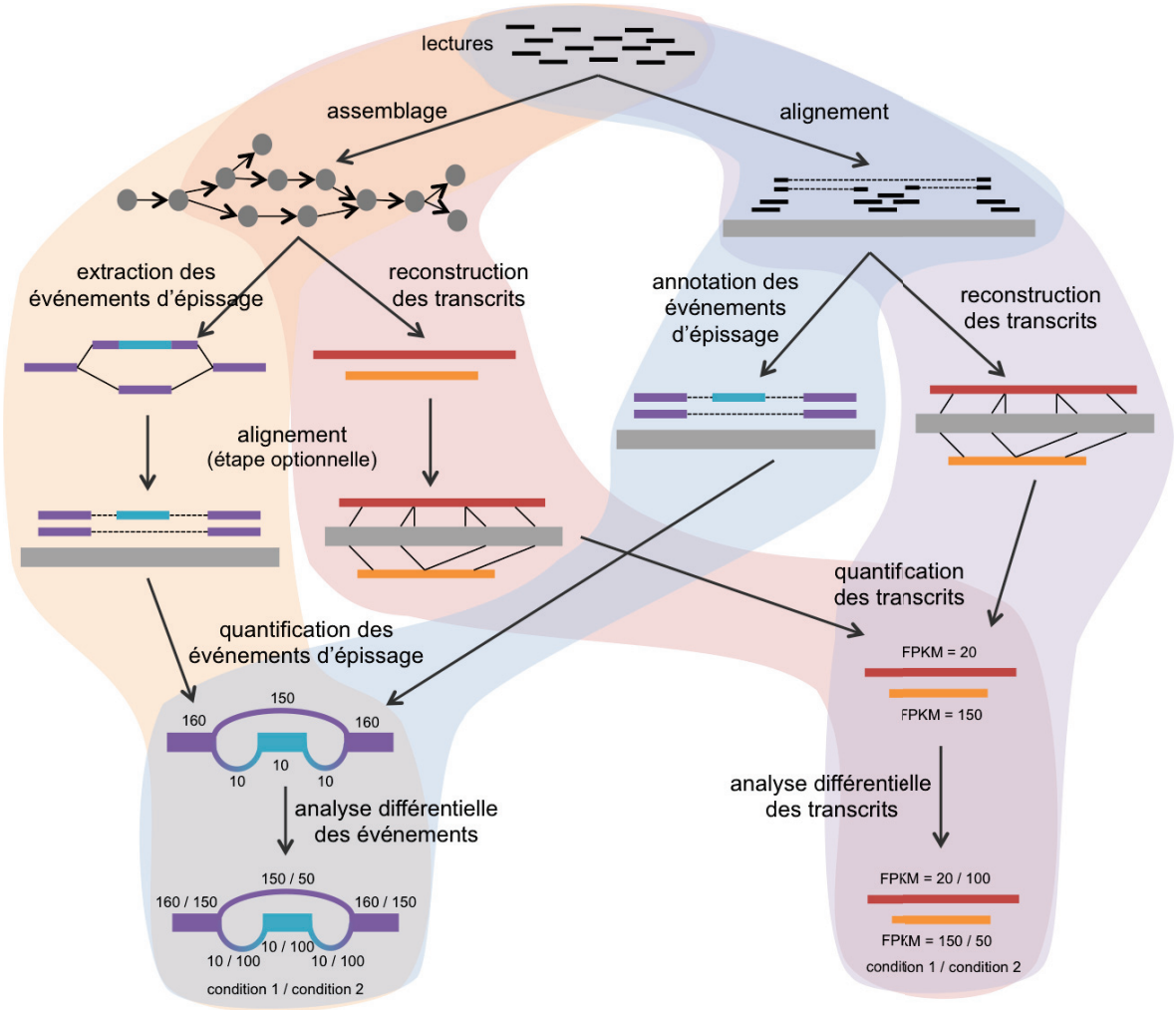


FIGURE 25 – **Pipeline d'analyse de l'épissage dans les données RNA-Seq.** Pour analyser l'épissage alternatif dans des données RNA-Seq, le premier choix à faire est entre les méthodes basées sur l'assemblage (partie gauche du schéma) et les méthodes basées sur l'alignement au génome de référence (partie droite du schéma). Ensuite, quel que soit le choix fait précédemment, il faudra choisir entre des méthodes locales (qui vont uniquement extraire les événements d'épissage) et des méthodes globales (qui vont reconstruire des transcrits complets). Pour les méthodes basées sur l'assemblage, les événements ou les transcrits assemblés peuvent être alignés sur un génome de référence. L'étape suivante est la quantification. Pour les méthodes globales, c'est l'expression des transcrits complets qui va être calculée et ce sont ces expressions qui vont être comparées entre différentes conditions. Alors que pour les méthodes locales, ce sont uniquement les lectures qui chevauchent l'événement d'épissage qui vont être comptabilisées et utilisées pour réaliser l'analyse différentielle.

qualité et qu'elles ne contiennent pas de biais important pouvant influencer le reste de l'analyse. FastQC est le plus connu. Il permet d'obtenir un rapport complet sur un jeu de données, avec notamment la qualité moyenne des lectures, la qualité par base, le pourcentage moyen en GC des lectures, le pourcentage par base de chaque nucléotide, la distribution de la longueur des lectures ou encore les séquences sur-représentées dans l'échantillon. Ce type de contrôle peut permettre d'identifier d'éventuelles contaminations avec du matériel d'autres organismes.

Ensuite, suivant les résultats, les lectures pourront être filtrées et/ou "trimmées" (troncature de la fin des lectures qui est parfois de mauvaise qualité). Ces opérations de nettoyage peuvent être réalisées avec des outils comme prinseq [Schmieder and Edwards, 2011] ou trimmomatic [Bolger et al., 2014]. Il est parfois nécessaire d'utiliser des outils pour éliminer des séquences adaptatrices qui pourraient être encore présentes dans les séquences. En effet, si ces portions de séquences ne sont pas retirées, une grande majorité des aligneurs ne parviendront pas à réaliser l'alignement et ces lectures seront alors jetées. Cette étape d'élimination des adaptateurs est donc très importante pour conserver le maximum d'information. Elle peut être réalisée avec des outils comme trimmomatic ou cutadapt [Martin, 2011].
Enfin, il est conseillé de re-vérifier la qualité des données après toutes ces étapes de nettoyage avant de se lancer dans la suite de l'analyse.

3 Alignement

Après obtention de données de bonne qualité, nous pouvons passer à l'étape suivante qui consiste à aligner les lectures à un génome de référence ou bien à assembler ces lectures, suivant le choix méthodologique fait. Dans cette section nous allons nous intéresser aux différents outils permettant de réaliser les alignements.
La quantification des isoformes et des événements d'épissage dépend grandement d'un alignement correct sur le génome de référence. Pour aligner des lectures provenant d'une expérience de RNA-Seq, il est nécessaire d'avoir une méthode qui peut aligner les lectures en plusieurs blocs (appelé *spliced mapper*). Les premières méthodes permettant de réaliser ce type d'alignement utilisaient une heuristique rapide pour associer les séquences avec un modèle pour les sites d'épissage. Parmi ces méthodes, on trouve BLAT [Kent, 2002] et GMAP [Wu and Watanabe, 2005]. Mais ces méthodes ne sont pas assez puissantes pour traiter la quantité de lectures générées par les séquenceur haut-débit en un temps raisonnable.
Une quantité de nouvelles approches ont été développées pour aligner les courtes séquences sur un génome de référence. La difficulté de l'alignement de lectures provenant du RNA-Seq se trouve dans le fait que dès lors que les lectures recouvrent une jonction d'exons, elles doivent être alignées en plusieurs blocs. Cela implique que ces lectures déjà courtes soient découpées en séquences entre plus petites rendant l'alignement de manière unique encore plus compliqué. Et même si les introns sont définis par des sites d'épissage spécifiques (les sites d'épissage canoniques sont GT-AG, GC-AG et AT-AC), ces séquences ne sont pas rares dans le génome et peuvent être

dues à la chance et ne pas définir un site d'épissage.

On classifie traditionnellement les *spliced mapper* en deux grandes catégories : les outils basés sur l'alignement sur l'exon (méthodes *exon-first*) et les outils qualifiés de *seed-and-extend*. Dans la première classe d'outils, les lectures sont alignées en un seul bloc au génome de référence, définissant ainsi des clusters de lectures (correspondant à des exons). Les lectures non alignées sont alors utilisées pour trouver des jonctions entre les clusters de lectures. Parmi ces méthodes, on trouve TopHat [Trapnell et al., 2009], SOAPsplice [Huang et al., 2011], PASSion [Zhang et al., 2012], MapSplice [Wang et al., 2010], SpliceMap [Au et al., 2010] et GEM [Marco-Sola et al., 2012]. La deuxième catégorie d'outils aligne d'abord un morceau de lecture en un seul bloc, puis étend cet alignement avec différents algorithmes. C'est au cours de cette étape que les sites d'épissage sont localisés. Parmi ces méthodes, on peut citer SplitSeek [Ameur et al., 2010], Supersplat [Bryant et al., 2010], SeqSaw [Wang et al., 2011], ABMapper [Lou et al., 2011], MapNext [Bao et al., 2009], STAR [Dobin et al., 2013], GSNAP [Wu and Nacu, 2010] et HISAT [Kim et al., 2015]. Ce type de méthode trouve généralement plus de nouvelles jonctions d'épissage. Mais certains de ces outils nécessitent beaucoup d'espace mémoire. Par exemple, pour réaliser un alignement sur le génome humain, STAR va nécessiter environ 30GB de RAM et il va réaliser l'alignement beaucoup plus rapidement que TopHat. HISAT a été développé très récemment et pourra probablement supplanter les autres outils grâce à sa vitesse et sa faible empreinte mémoire comparée à STAR pour une précision comparable. Enfin, il existe encore d'autres outils qui ne rentrent dans aucune de ces deux catégories car ils utilisent l'annotation ou certaines heuristiques pour réaliser l'alignement. Parmi ces outils, on peut citer RUM [Grant et al., 2011], SpliceSeq [Ryan et al., 2012] et PASTA [Tang and Riva, 2013].

D'autres critères sont également important à prendre en compte. Certains outils alignent les lectures en utilisant les annotations comme guide. Les alignements sur les jonctions connues sont alors plus précis. Mais pour pouvoir trouver de nouvelles jonctions, les annotations ne doivent pas être contraignantes. Par exemple, TopHat peut prendre en entrée une annotation pour guider l'alignement mais les lectures non alignées grâce à ces annotations sont ensuite utilisées pour trouver de nouvelles jonctions. D'autres paramètres peuvent être importants, comme la longueur minimum et maximum des introns ou les sites d'épissage pouvant être utilisés (uniquement les canoniques, ou possibilité d'aligner des lectures sur des sites non-canoniques).

L'assignation des lectures à une région unique du génome peut être difficile pour les méthodes d'alignement et poser des problèmes pour la suite de l'analyse. En effet, il est difficile de savoir

quoi faire avec les lectures qui s'alignent de manière exacte à plusieurs endroits du génome. Le plus souvent ces lectures sont mises de côté et ne sont donc pas utilisées pour la suite de l'analyse. Cela entraîner une perte de puissance pour les analyses statistiques de certaines régions du génome (comme les éléments répétés). Le choix final de l'outil d'alignement dépend donc à la fois des moyens matériels et si le but est d'aligner les lectures sur des jonctions connues ou de trouver de nouvelles jonctions.

4 Assemblage

L'alternative à l'alignement est l'assemblage. Cette méthode est particulièrement utilisée dans le cas où il n'y a pas de génome de référence ou bien lorsque le génome des individus est trop différent du génome de référence (comme cela peut être le cas dans le cancer). Mais cette méthodologie peut également être utilisée dans le cas où un génome de référence est disponible. Dans ce cas là, les transcrits ou les événements d'épissage assemblés peuvent être alignés sur le génome. Ces séquences étant plus longues que les lectures d'origines, il est plus aisé de les assigner de manière unique à une région du génome.

Les assembleurs se servent des chevauchements entre lectures pour les assembler. Toutes les séquences de longueur k (appelés k-mers) sont extraites des lectures. Ces k-mers vont permettre de construire un graphe de *de Bruijn*. Cette structure de données est un graphe orienté qui permet de représenter tous les chevauchements de longueur k-1 entre tous les k-mers. Les noeuds de ce graphe sont donc les k-mers et les arrêtes relient les noeuds qui ont des séquences qui se chevauchent de k-1 nucléotides. La valeur de k est une variable qui va impacter le résultat de l'assemblage : plus grande sensibilité avec de petites valeurs de k et plus grande spécificité avec des grandes valeurs de k. Cette structure de données permet ensuite d'extraire les transcrits pour les outils tels que Rnnotator [Martin et al., 2010], OASES [Schulz et al., 2012], Trans-ABySS [Robertson et al., 2010] et Trinity [Grabherr et al., 2011] ou bien les événements d'épissage pour les outils tels que KisSplice [Sacomoto et al., 2012].

Ce type de méthodes est encore peu utilisé dans les cas où un génome de référence est disponible car elles sont réputées pour nécessiter beaucoup de RAM. Mais il y a eu des progrès récents dans les méthodes de stockage en mémoire des graphes de *de Bruijn*, permettant ainsi de réduire l'empreinte mémoire de ce type de méthode. Le principal avantage de cette approche réside dans le fait qu'elle ne dépend pas de l'alignement des lectures à un génome de référence. Les nouveaux exons ou sites d'épissage sont donc traités comme les exons ou les sites connus. De plus, la taille

des introns n'est pas une limite non plus pour les assembleurs. Pas contre, ces méthodes nécessitent souvent une profondeur de séquençage plus importante que les méthodes d'alignement pour reconstruire les transcrits complets ou annoter les événements d'épissage. Les assembleurs sont également plus sensibles aux erreurs de séquençage.
Ce type de méthode est encore peu utilisé pour l'analyse de l'épissage, alors qu'il est clairement avantageux dans l'annotation de nouveaux exons ou événements d'épissage.

5 Identification des événements d'épissage ou des isoformes

L'identification des événements d'épissage peut être fait à partir des annotations pour les méthodes comme MISO [Katz et al., 2010] basées sur l'alignement des lecture. Les événements extraits à partir des annotations sont alors classés par type d'événements (exon cassette, donneur ou accepteur alternatif, exons mutuellement exclusif, ...).
Mais l'annotation des événements d'épissage peut également être fait à partir de l'assemblage des lectures. C'est ce que l'outil KisSplice fait. Les événements d'épissage forment des structures spécifiques dans le graphe de *de Bruijn* appelées "bulles". Les 2 chemins de la bulle représentent les 2 variants d'épissage. La classification des événements par type est réalisé après alignement des chemins sur le génome de référence.

La reconstruction des isoformes aussi peut être faite à partir de l'alignement ou de l'assemblage des lectures. Parmi les outils permettant de reconstruire des isoformes à partir de l'alignement des lectures, on trouve Cufflinks [Trapnell et al., 2010], Scripture [Guttman et al., 2010], StringTie [Pertea et al., 2015] et FlipFlop [Bernard et al., 2014]. La première étape de Cufflinks est de reconstruire un graphe de chevauchement ("overlap graph") à partir de toutes les lectures qui s'alignent sur un locus du génome. Puis, ce graphe est parcouru pour reconstruire les isoformes en considérant le plus petit ensemble de transcrits qui permet d'expliquer les lectures (principe de parcimonie). Scripture construit un graphe d'épissage. Dans ce type de graphe, les noeuds représentent des exons ou des morceaux d'exons et les arêtes des variations d'épissage. Ensuite, Scripture récupère tous les chemins du graphe qui ont une couverture significative. Cet ensemble de chemin définit les transcrits. StringTie utilise comme Scripture un graphe d'épissage. Mais en plus, StringTie prend en compte l'abondance des transcrits pour faire la reconstruction. La reconstruction et la quantification des transcrits sont donc faites en même temps. FlipFlop, comme StringTie, prend en compte l'abondance des transcrits pour réaliser leur reconstruction. Ce type

de méthode prenant en compte la quantification afin de reconstruire les transcrits, sont plus précises que les autres méthodes.

Enfin, la reconstruction des transcrits peut également être réalisée à partir de l'assemblage. Trinity, OASES et Trans-ABySS sont des exemples d'outils permettant d'obtenir des transcrits en réalisant un assemblage des lectures. Ces outils parcourent le graphe de *de Bruijn* construit à l'étape d'avant pour pouvoir en déduire les transcrits exprimés dans l'échantillon étudié.

6 Quantification

L'étape de quantification consiste à compter le nombre de lectures supportant les différentes isoformes d'un gène ou bien plus localement supportant les différents variants d'un événement d'épissage. Cette étape est souvent incluse dans l'outil d'identification des événements d'épissage ou de reconstruction des transcrits.

Les outils de quantification de l'expression des transcrits (comme Cufflinks, Scripture ou StringTie) vont assigner les lectures à chaque isoforme pour pouvoir réaliser la quantification (Figure 26 B). Ils vont calculer un RPKM (Reads per kilobase per million mapped reads) ou FPKM (Fragment per kilobase per million mapped reads). Cette mesure donne une valeur d'expression relative comparable entre différentes conditions et entre différents gènes.

Les outils locaux, comme KisSplice, utilisent les lectures totalement inclus dans les exons d'intérêt et/ou les lectures chevauchant les jonctions d'exons (Figure 26 A). Ces lectures jonctions sont les plus spécifiques de chaque variant d'épissage alors que les lectures exoniques s'alignant dans les exons flanquants l'exon alternatif peuvent provenir tout aussi bien de la forme incluse que de la forme exclue. La mesure retournée par de nombreux outils locaux est le pourcentage d'inclusion, noté PSI (percent spliced in). D'autres outils locaux, comme DEXSeq, quantifient chaque exon ou morceau d'exon séparément et n'utilisent pas l'information des lectures jonctions. Enfin, certains outils comme MISO peuvent faire soit une quantification des isoformes, soit des événements d'épissage suivant le choix de l'utilisateur.

7 Analyse différentielle

L'objectif de l'analyse différentielle est de tester si l'épissage est modifié entre deux conditions. L'analyse peut être faite sur les transcrits complets avec des outils comme Cuffdiff2 [Trapnell et al., 2013] et Ballgown (librairie R disponible dans Bioconductor) ou plus localement sur des événements

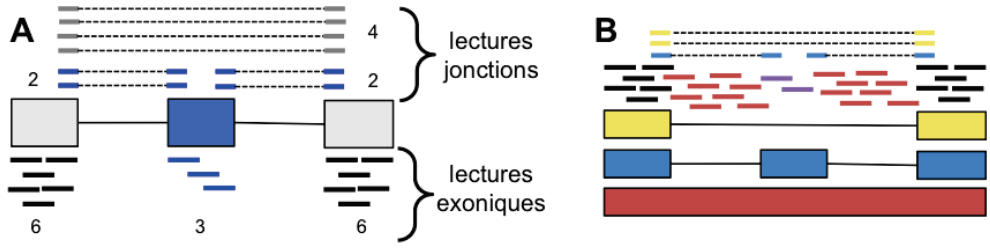


FIGURE 26 – **Analyse de l'épissage : la quantification. A.** La quantification d'un événement d'épissage (ici un exon cassette) peut être faite en comptant le nombre de lectures jonctions (lectures chevauchant au moins deux exons), le nombre de lectures exoniques ou les deux types de lectures. **B.** Lors de la quantification des isoformes reconstruites, les lectures sont assignées aux transcrits dont elles peuvent provenir. Les lectures grises peuvent provenir des 3 transcrits et les lectures violettes du transcrit bleu ou rouge. La majorité des outils utilise des méthodes statistiques pour réaliser l'assignation des lectures à un transcrit et ensuite elles en déduisent l'expression des différentes isoformes.

d'épissage ou les exons avec des outils comme MISO, MATS [Shen et al., 2012] et DEXSeq [Anders et al., 2012].

DEXSeq utilise le nombre de lectures par exons pour calculer les gènes avec un épissage différentiel entre deux conditions. Par contre, il ne donne pas d'information sur les événements, seulement sur les exons ou morceaux d'exons dont l'expression est modifiée. MISO et MATS analysent les événements d'épissage par type et reportent les événements différentiellement régulés entre les deux conditions comparées. Ils utilisent tous les deux une approche bayésienne pour calculer le différentiel d'inclusion de l'événement entre les deux conditions.

Un point important à prendre en compte lors d'une analyse différentielle est la variabilité biologique. Pour cela, on utilise des réplicats biologiques. Or certaines méthodes comme MISO et la première version de MATS ne prennent pas en compte cette variabilité biologique. DEXSeq et rMATS [Shen et al., 2014] une nouvelle version de MATS acceptent les réplicats biologiques et modélisent la variabilité biologique lors de l'analyse.

Cuffdiff2 et Ballgown ne permettent pas d'étudier les événements d'épissage localement, mais seulement d'étudier l'expression des différentes isoformes d'un gène. Ils peuvent tous les deux prendre en compte les réplicats biologiques.

8 Visualisation

Une première manière pour visualiser les données RNA-Seq consiste à transformer les données dans un format chargeable sur le visualiseur de génome d'UCSC disponible en ligne. On peut alors parcourir les données et les comparer avec les annotations présentes dans UCSC. Mais lorsque

les données sont trop volumineuses, il est parfois compliqué de les charger dans ce visualiseur. Certains outils ont été développés spécifiquement pour visualiser les données haut-débit en local sur nos propres machines. Le plus connu et le plus utilisé de ces outils est IGV (Integrative Genome Viewer).

Dans IGV, la visualisation des données RNA-Seq est composée de 3 parties : la couverture des lectures sur le génome, les jonctions et la visualisation de toutes les lectures individuellement (Figure 27 A). Cet outil permet également de réaliser des "sashimi plot" (Figure 27 B). Ce type de visualisation permet de représenter sur la même figure la couverture et les jonctions. De plus, les jonctions sont quantifiées.

La figure 27 C montre le zoom sur un événement de cassette exon. On voit que des lectures supportent à la fois l'inclusion et l'exclusion de cet exon du gène ENAH. De plus, on peut voir que la forme excluant l'exon est largement plus exprimée que la forme l'incluant dans l'échantillon visualisé. Ceci est visible grâce à la quantification des lectures jonctions (1 082 lectures supportant l'exclusion et 19 et 23 lectures supportant l'inclusion), mais également grâce à la couverture des lectures, qui montre une très faible couverture de cet exon comparé à ses exons flanquants.

Certains outils d'analyse de l'épissage incluent directement un module de visualisation, comme DEXSeq. Cette librairie R inclue une fonction permettant de représenter le niveau d'expression de chaque exon ou morceau d'exon pour les différentes conditions comparées.

9 Pour aller plus loin

Une fois que l'on connaît les événements d'épissage qui sont différentiellement régulés ou les isoformes exprimées dans une condition d'intérêt, on peut analyser les candidats un par un ou essayer de les analyser de manière plus globale pour comprendre l'impact de l'ensemble de ces variations sur le fonctionnement de la cellule.

Pour cela, il existe des outils permettant d'intégrer différents types de données haut-débit (par exemple, RNA-Seq et ChIP-Seq), de déterminer les gènes co-régulés au sein d'un réseau ou d'analyser des voies de signalisation affectées par des dérégulations.

Mais la plupart de ces analyses intègrent principalement le niveau d'expression des gènes et pas l'épissage. D'où la nécessité de développer des outils adaptés pour interpréter les conséquences de variations d'épissage et aller plus loin dans la compréhension des conséquences de la régulation de ce processus.

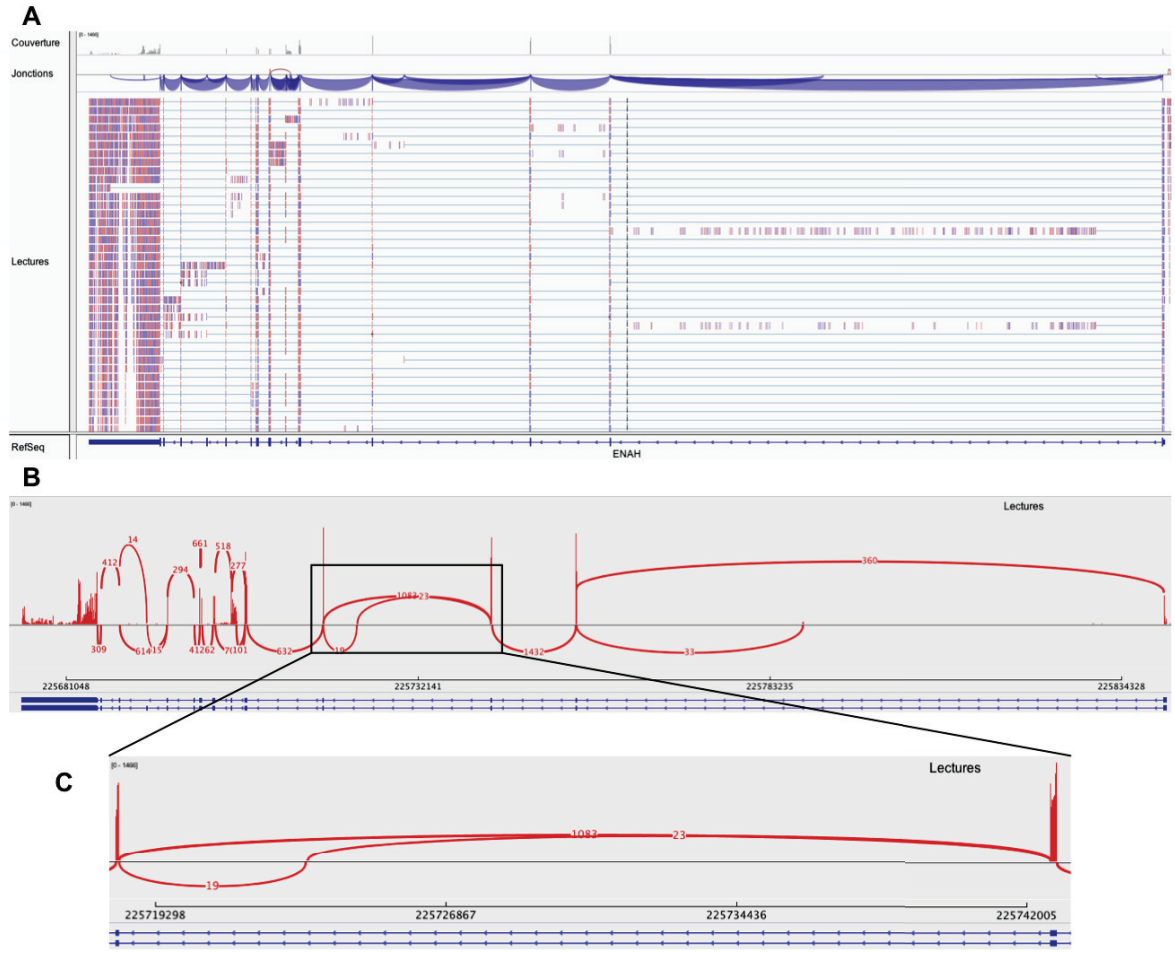


FIGURE 27 – **La visualisation des données RNA-Seq A.** Visualisation par défaut d'IGV. La première partie montre la couverture des lectures sur le gène ENAH. Ensuite, on peut voir les jonctions représentées par des arcs de cercles plus ou moins épais suivant la quantité de lecture qui supporte une jonction. Puis, on peut visualiser toutes les lectures individuellement. La dernière partie représente l'annotation du gène ENAH d'après RefSeq. **B.** La visualisation "sashimi plot" permet de représenter sur une même figure la couverture et les jonctions. De plus, le nombre de lectures supportant chaque jonction est affiché sur chaque ligne représentant une jonction. L'exemple affiché correspond au gène ENAH. **C.** Zoom du "sashimi plot" sur un événement de cassette exon. L'exclusion de l'exon est supportée par 1 082 lectures et son inclusion par 19 à gauche et 23 à droite. En regardant la couverture des lectures sur cette portions du génome, on peut confirmer que cet exon est moins présent dans les transcrits que ses exons flanquants.

Chapitre 2

Objectifs

Avant mon arrivée dans l'équipe, l'analyse de l'épissage alternatif était réalisée grâce à des puces à ADN. Une plateforme web, nommée Elexir, avait été développée pour réaliser de type d'analyse. Mon rôle au sein de l'équipe fut de mettre en place des méthodes permettant d'étudier les données de séquençage à haut débit et notamment l'épissage alternatif dans des données RNA-Seq.

Nous voulions une méthode d'analyse de l'épissage alternatif qui réponde à un cahier des charges précis. Tout d'abord, nous voulions pouvoir analyser différents types d'événement d'épissage (exon cassette, site donneur ou accepteur alternatif, ...). De plus, nous souhaitions que la méthode puisse utiliser des lectures de longueur variable, comme celles issues du séquenceur Roche 454. En effet, nous avions un séquenceur Roche 454 Junior à disponibilité au laboratoire. Ensuite, l'approche devait être capable de prendre en compte les lectures jonctions (lectures chevauchant au moins 2 exons) dans l'étape de quantification. Et elle devait être capable de prendre en compte les réplicats biologiques dans l'analyse différentielle. Enfin, nous voulions une méthode flexible par rapport aux annotations afin de ne pas rater trop d'événements à cause d'annotations incomplètes.

Il existe un grand nombre de méthodes permettant d'analyser l'épissage alternatif dans les données RNA-Seq mais aucune ne correspondait à tous nos critères. Le premier objectif de ma thèse a donc été de développer un outil d'analyse de l'épissage alternatif dans les données RNA-Seq répondant à notre cahier des charges. Cette approche a été nommé FaRLine pour FasterDB RNA-seq Pipeline.

Pour permettre une analyse facile et rapide des données RNA-Seq aussi bien au niveau de l'épissage alternatif et au niveau de l'expression des gènes, j'ai automatisé le pipeline d'analyse sur un cluster de calcul.

J'ai utilisé ces pipelines pour analyser de nombreux jeux de données pour des projets de l'équipe ou pour des collaborateurs. Dans le cas où de nouvelles données étaient été générées, j'ai participé à la discussion avec les biologistes sur le plan d'expérience du RNA-Seq. Ces discussions entre biologiste et bioinformaticiens sont particulièrement importantes pour générer des données exploitables et qui permettent de répondre à la question biologique. Notamment, ces discussions ont permis aux biologistes de l'équipe de prendre conscience de l'importance des réplicats biologiques pour obtenir des résultats fiables.

Dans un deuxième temps, nous avons comparé FaRLine à KisSplice, une méthode développée par l'équipe avec laquelle nous collaborons. Notre approche est basée sur l'alignement des lectures sur un génome de référence, alors que KisSplice est basé sur l'assemblage des lectures. L'objectif de ce projet était d'évaluer les différences en terme d'annotation et de quantification de ces deux approches, afin d'en déduire leurs avantages et leurs inconvénients.

Mais connaître les exons ou morceaux d'exons inclus dans une condition physiologique donnée n'est pas le but final pour les biologistes. Il faut encore interpréter les conséquences des variations d'épissage au niveau des protéines et des cellules. Pour cela, j'ai participé au développement d'outils pour aller plus loin dans l'analyse de l'épissage et comprendre l'impact fonctionnel de modifications d'épissage.

Chapitre 3

Résultats

I Développement d'un pipeline d'analyse de l'épissage alternatif dans les données RNA-Seq

La première partie de mon travail de thèse a consisté à mettre en place un outil d'analyse de données RNA-Seq qui permette d'étudier l'épissage alternatif et l'expression des gènes. L'objectif était que cet outil soit utilisable facilement par les bioinformaticiens de l'équipe et qu'il respecte toutes les caractéristiques du cahier des charges comme décrit dans la partie Objectif.

A FasterDB

L'outil d'analyse des données RNA-Seq a été conçu autour de FasterDB. FasterDB est la base de donnée de l'équipe qui répertorie les transcrits de l'homme et de la souris [Mallinjoud et al., 2014]. Cette base de données est liée à un outil de visualisation web.
Pendant ma thèse, j'ai mis à jour la partie de cette base de données consacrée à l'homme. En effet, la version de FasterDB qui était utilisée était basée sur la version 60 d'EnsEMBL et ne contenait que les gènes codants. La mise à jour a été réalisée à partir des annotations de la version 75 d'EnsEMBL (qui est la dernière version des annotations EnsEMBL sur le génome hg19). Les informations concernant les gènes, les transcrits et les exons de transcrits ont été récupérés via l'interface de programmation (API ou "Application Programming Interface") d'EnsEMBL. Ensuite, j'ai utilisé les scripts de FasterDB pour définir les exons génomiques et les annotations des différents événements d'épissage (exon cassette, accepteur et donneur alternatif, intron rétention). La génération des exons génomiques est réalisée en projetant les exons de transcrits tout en essayant le plus possible de maximiser le nombre d'exons. Les bornes des exons génomiques sont les bornes définissant le plus grand exon génomique sans prendre en compte les premiers et dernier exons alternatifs (Figure 28). Lorsque certains transcrits définissent un seul exon et d'autres en définissent plusieurs sur la même région génomique, ce qui est le plus représenté dans les transcrits est pris comme référence.

Tous les outils décrits par la suite utilisent cette base de données comme référence, à la fois pour les gènes et pour les événements d'épissage.

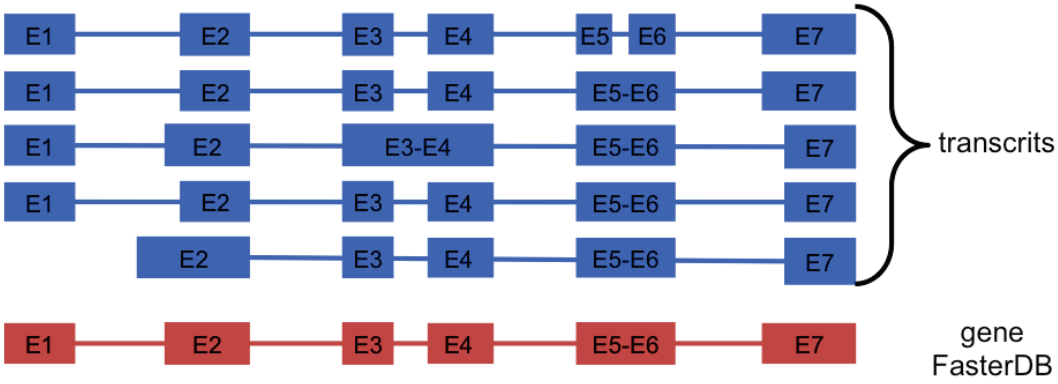


FIGURE 28 – **Définition des exons génomiques dans FasterDB.** Les exons génomiques sont définis à partir des exons des transcrits. Les bornes choisies sont celles qui définissent l'exon génomique le plus grand sans prendre en compte les premiers ou derniers exons alternatifs.

B FaRLine

FasterDB RNA-seq Pipeline, FaRLine, est l'outil d'analyse de l'épissage alternatif que j'ai développé pour l'équipe avec Emilie Chautard, ancienne post-doctorante au sein de l'équipe.
Ce pipeline comprend plusieurs étapes décrites sur la Figure 29, dont certaines sont réalisées avec des outils extérieurs. Notamment, la première étape qui est l'alignement des lectures sur le génome de référence, est réalisé avec TopHat2 dans notre pipeline. Mais elle peut également être réalisée avec d'autres outils d'alignement, comme par exemple STAR. Il est préférable de réaliser l'alignement avec des annotations comme guide. En effet, cela permet d'avoir de meilleurs alignements notamment au niveau des éléments répétés, comme les ALU.

L'étape suivante est l'annotation des événements d'épissage à partir de l'alignement des lectures. Toutes les lectures s'alignant en plusieurs blocs sont parcourues. Les lectures jonctions qui excluent un ou plusieurs exons permettent de définir les événements de saut d'exons (simple ou multiples) et les jonctions qui ne sont pas alignées sur un site d'épissage connu permettent de définir de nouveaux sites accepteurs ou donneurs.
Puis, ces événements sont quantifiés. Notre outil prend en compte uniquement les lectures jonctions dans cette étape de quantification. Ces comptages nous permettent de calculer un pourcentage d'inclusion de l'événement, noté Ψ (PSI signifiant "Percent Spliced In"). Le Ψ est calculé en faisant un ratio entre le nombre de lectures supportant l'inclusion de l'événement et le nombre total de lectures supportant l'événement (inclusion et exclusion). Un exemple de calcul de Ψ est donnée dans la figure 30.

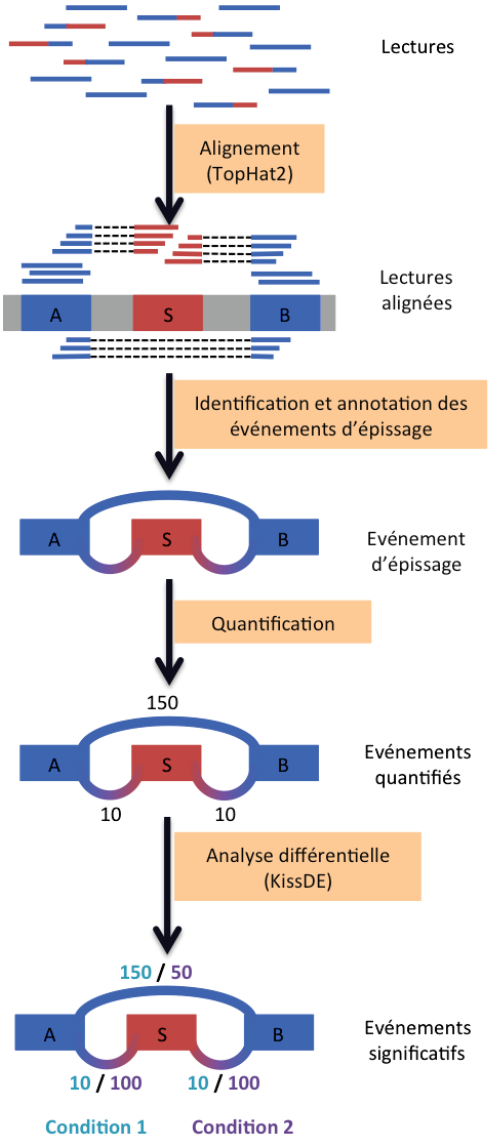


FIGURE 29 – **Le pipeline d'analyse de l'épissage : FaRLine.** Les lectures sont d'abord alignées sur le génome de référence avec TopHat2. Les lectures alignées et les annotations sont ensuite utilisées pour annoter et classifier les différents types d'événement d'épissage. Puis chaque événement d'épissage est quantifié. Lors de cette quantification, seules les lectures jonctions sont prises en compte. Après avoir réalisé ces différentes étapes pour chaque échantillon, on peut réaliser l'analyse différentielle. Les conditions sont comparées et on extrait ainsi les événements qui sont significativement régulés entre chaque condition.

Les différentes catégories d'événements sont analysées séparément. Par exemple, lors de l'analyse d'un saut d'exon, si l'exon flanquant l'exon alternatif possède deux sites donneurs possibles, alors nous allons compter toutes les lectures qui supportent l'exclusion quelques soient leurs sites donneurs (Figure 30).

Certains événements peuvent être réassignés à une autre catégorie d'événement en fonction de la quantification. En effet, un événement de saut de deux exons peut être redéfini en un exon cassette si lors de la quantification, on se rend compte qu'un des deux exons n'est supporté par

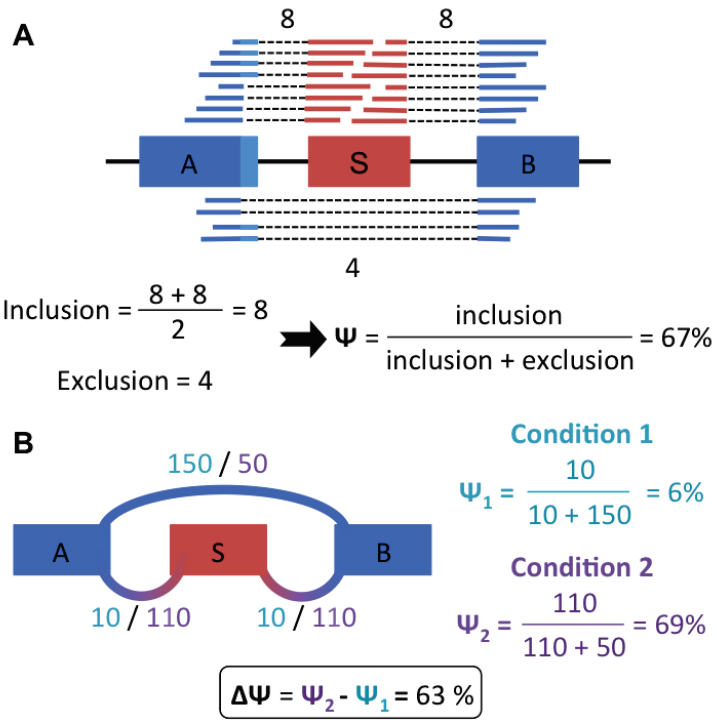


FIGURE 30 – **Calcul du Ψ et du ΔΨ pour un événement d'exon cassette. A.** Lors de l'étape de quantification, les lectures jonctions supportant l'événement sont dénombrées. Dans le cas de l'analyse d'un exon cassette avec un site donneur alternatif sur l'exon flanquant A (en amont), les lectures jonctions correspondant aux différents sites d'épissage sont comptabilisées. Dans l'exemple, il y a donc 8 lectures supportant la jonction AS, 8 supportant la jonction SB et 4 supportant la jonction AB. Le nombre de lectures d'inclusion est calculé en faisant la moyenne du nombre de lectures supportant les deux jonctions d'inclusion (AS et SB). On obtient le Ψ en calculant la ratio du nombre de lecture d'inclusion sur le nombre de lecture d'exclusion. Dans notre exemple, $\Psi = 8/(8+4) = 67\%$. **B.** Lors de l'analyse différentielle, le Ψ de chacune des conditions est calculé. Ensuite, on fait la différence de ces deux Ψ pour obtenir le ΔΨ. Dans l'exemple représenté, le Ψ de la première condition vaut 6%, celui de la deuxième condition vaut 67%. Le ΔΨ vaut donc 61%.

aucune lecture jonction partant des autres exons contenus dans l'événement (exons flanquants et exons alternatifs). Un exemple de ce type de cas est donné dans la figure 31 A. C'est également à cette étape que l'on définit les événements d'exons mutuellement exclusifs à partir des exons annotés comme exon cassette simple (Figure 31 B). Pour que deux exons soient définis comme mutuellement exclusifs, il faut que le nombre de lectures qui supportent la jonction entre ces deux exons soit bien inférieur au nombre de lectures supportant les autres jonctions de l'événement. Le seuil a été empiriquement fixé à 5%. C'est-à-dire que le nombre de lectures supportant la jonction entre les deux exons doit être inférieur ou égal à 5% du nombre de lectures de la jonction la moins exprimée de l'événement.

Enfin, nous pouvons réaliser l'analyse différentielle. Lorsque les conditions que l'on souhaite com-

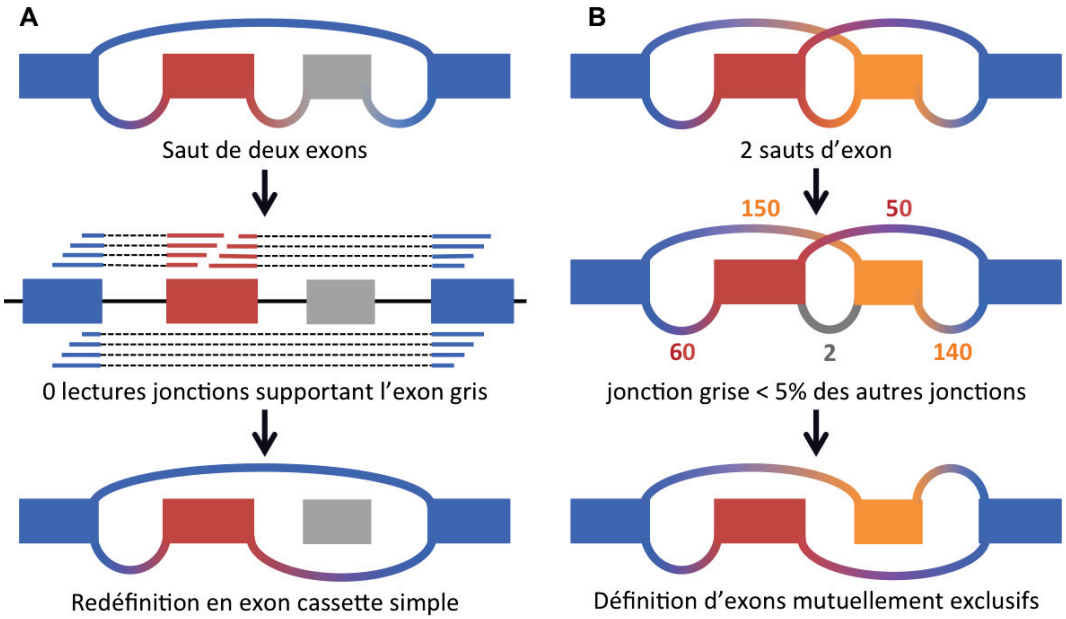


FIGURE 31 – **Redéfinition de certains types d'événement d'épissage. A.** Les sauts de multiples exons sont définis lorsqu'une lecture jonction exclut au moins deux exons. Lors de l'étape de quantification, si un des exons alternatifs de cet événement n'est supporté par aucune lecture jonction, on l'enlève de l'événement. Le saut de multiples exons est alors redéfinit comme un exon cassette simple. **B.** Deux événements d'exon cassette qui se suivent peuvent être redéfinit comme exons mutuellement exclusifs. C'est le cas lorsque la jonction reliant les deux exons alternatifs est très faiblement exprimée dans les données. Dans cet exemple, la jonction grise est comparée à la jonction de l'événement supportée par le moins de lectures. La jonction grise ne représente que 4% de la jonction supportée par 50 lectures. Cet événement est donc classé dans la catégorie des exons mutuellement exclusifs.

parer ne comportent pas de réplicat, on utilise un test de Fisher exact. Dans le cas où chaque condition comporte au moins deux réplicats biologiques, on utilise la librairie R kissDE. Cette librairie a été développée comme une suite de KisSplice par nos collaborateurs de l'équipe INRIA Baobab/Erable et elle est disponible à cette adresse : `http://kissplice.prabi.fr/tools/kissDE/`. De plus, cette librairie R devrait bientôt être disponible sur Bioconductor. J'ai participé à son développement et je l'ai intégrée dans notre pipeline. La méthode statistique utilisée par KissDE a été présentée et publiée dans l'article suivant [Lopez-Maestre et al., 2016].

Les événements considérés comme significativement différentiellement régulés entre les deux conditions sont ceux qui ont une p-value ajustée inférieure à 0.05 et un $\Delta\Psi$ supérieur à 0.10 en valeur absolue. Le $\Delta\Psi$ définit la variation de l'inclusion entre les deux conditions comparées. Il est calculé en réalisant la différence entre les Ψ des deux conditions (Figure 30). Plus le $\Delta\Psi$ va être grand (en valeur absolue), plus la variation entre les deux conditions est importante.

C Pipeline complet

Notre objectif n'était pas uniquement d'analyser l'épissage alternatif dans les données RNA-Seq. Nous souhaitions également pouvoir connaître le niveau d'expression des gènes. Nous avons donc développé un pipeline dédié à ce type d'analyse.

Pour cela, nous avons choisi d'utiliser des méthodes déjà publiées et bien validées dans la littérature. Les différentes étapes sont représentées sur la figure 32. Pour l'étape d'alignement, nous utilisons les fichiers générés par TopHat2 dans le pipeline d'analyse de l'épissage alternatif. Nous avons ensuite choisi HTSeq-count [Anders et al., 2015] pour réaliser la quantification des gènes et DESeq2 [Love et al., 2014] pour l'analyse différentielle. Ce pipeline ne marche que dans le cas où des réplicats biologiques sont disponibles. Sinon, l'analyse est réalisée avec l'outil GFold [Feng et al., 2012].

Les deux pipelines fournissent en sorti des fichiers excel faciles à manipuler par les biologistes. Une visualisation web est également disponible.

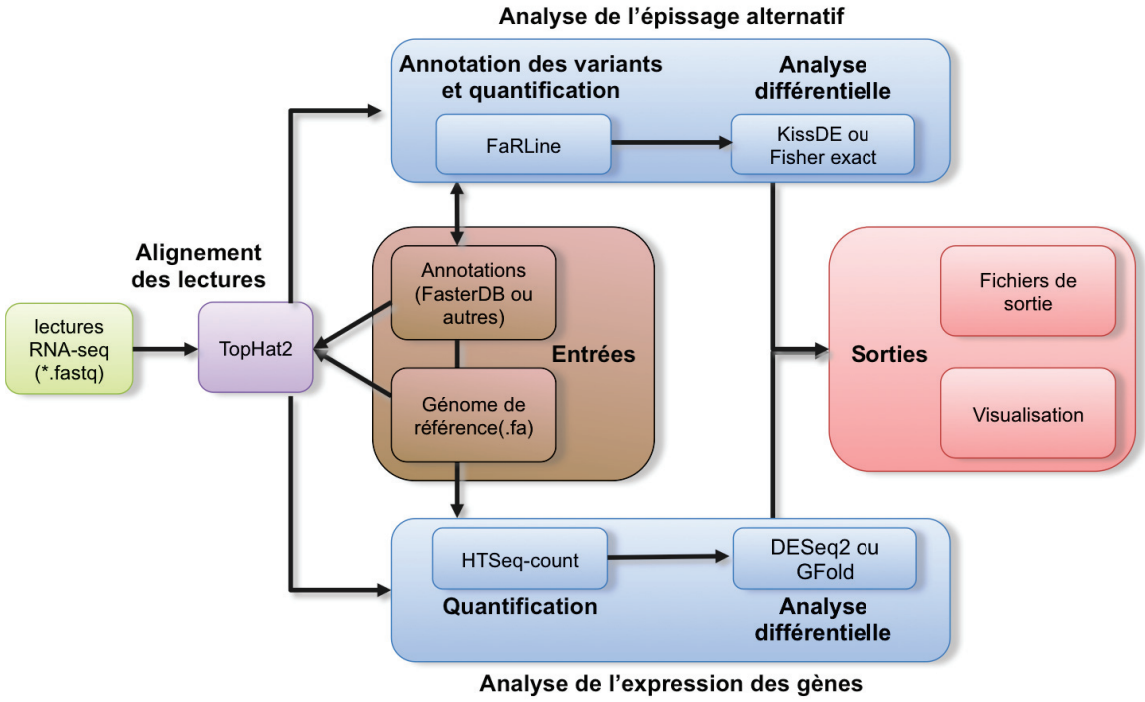


FIGURE 32 – **Le pipeline d'analyse de l'expression et de l'épissage.** Le pipeline complet est basé sur l'alignement des lectures sur le génome de référence. L'analyse de l'épissage est réalisée avec FaRLine comme expliqué précédemment. L'analyse de l'expression des gènes est réalisée avec des outils bien validés dans la littérature : HTSeq-count, DESeq2 et GFold. En sortie, les biologistes ont accès à des fichiers excel contenant tous les résultats (événements d'épissage significatifs et gènes différentiellement exprimés) ainsi qu'à une visualisation web.

Afin de pouvoir lancer des analyses de données RNA-Seq en routine, nous avons automatisé le

lancement des différentes étapes du pipeline d'analyse de l'épissage ainsi que celui d'analyse de l'expression des gènes.

Les deux pipelines ont été automatisés sur un cluster de calcul avec l'aide de Lisa Guigue, ancienne ingénieure d'étude dans l'équipe. Il suffit de remplir un fichier de configuration, puis de lancer le script bash d'automatisation pour que toutes les étapes soient lancées. Les étapes qui dépendent d'autres étapes sont mises en attentes jusqu'à la fin des étapes la précédant.

Les étapes de contrôle qualité des séquences en amont des analyses sont semi-automatisées. Elles peuvent être lancées sur tous les échantillons en une fois, mais elles nécessitent encore une intervention de l'utilisateur pour décider quels filtres sont nécessaires.

Aujourd'hui, ces pipelines sont utilisés en routine par les bioinformaticiens de l'équipe.

D Visualisation RNA-Seq

La visualisation RNA-Seq est basée sur celle de FasterDB. Elle a été principalement développée par Emilie Chautard.

Elle peut se décomposer en 3 grandes parties (Figure 33 A). La première partie de cette visualisation représente le gène défini dans FasterDB. C'est-à-dire que ce sont les exons génomiques qui sont affichés avec les jonctions dites canoniques (reliant deux exons génomiques successifs) en noir et les jonctions non canoniques en rouge.

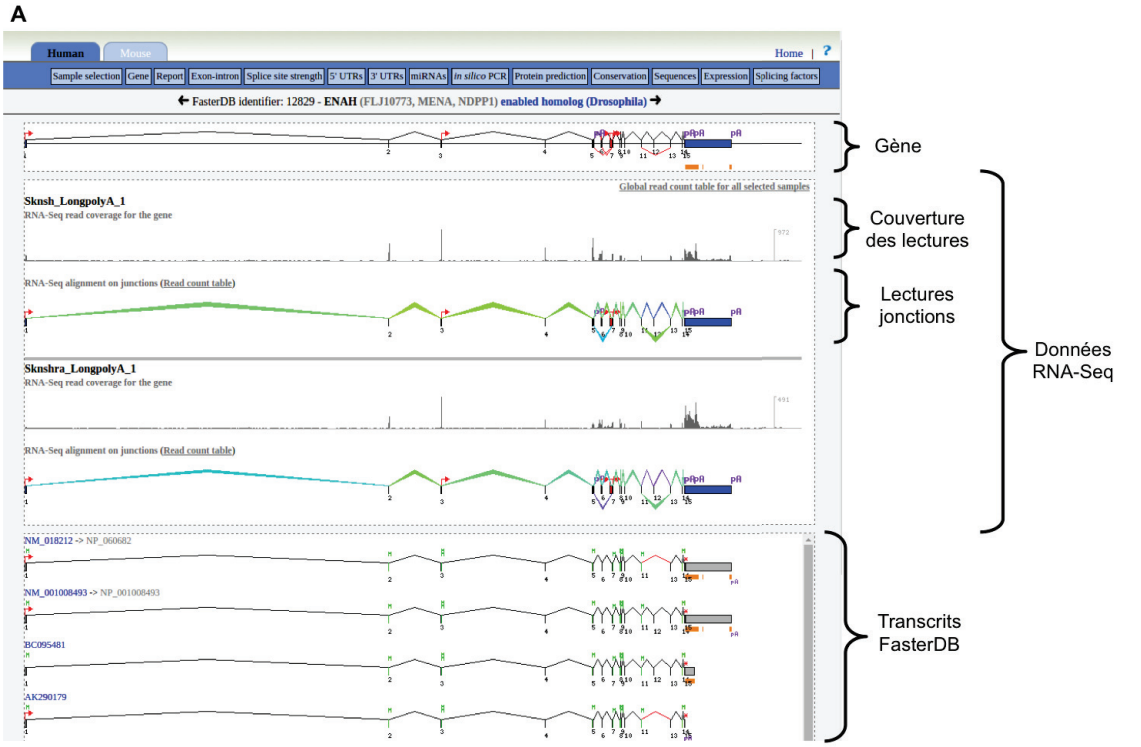
Ensuite, vient la visualisation des données RNA-Seq. Il est possible de choisir un ou plusieurs échantillons dans l'onglet "Sample selection". Pour chaque échantillon, on peut voir deux parties. La première représente la couverture des lectures sur le gène. La deuxième représente les lectures jonctions. Plus les traits sont épais, plus le nombre de lectures supportant la jonction est important. La couleur aussi est un indicateur du nombre de lectures supportant la jonction (Figure 33 C).

Enfin, comme dans la visualisation FasterDB d'origine, on peut voir les différents transcrits de ce gène.

Associé à la visualisation RNA-Seq, on peut afficher les quantifications des différentes jonctions représentées (Figure 33 B). Ce tableau contient à la fois les jonctions canoniques et non canoniques. Dans le cas d'un exon avec un site donneur ou accepteur alternatif, toutes les jonctions possibles seront présentes dans le tableau avec l'information de la distance du site alternatif par rapport à la borne de l'exon génomique défini dans FasterDB.

Cette interface permet de visualiser très rapidement les profils d'épissage dans différentes condi-

tions et de les comparer.



B

Junction read - raw count tables

Junction (A:B)	Gene coordinates	Chromosomal coordinates	Junction length	Donor (A)	Acceptor (B)	Counts Sknsh_LongpolyA_1	Counts Sknshra_LongpolyA_1
10:11	141334-145126	chr1:225695720-225699512	3793	none	none	294	250
11:12	145194-148090	chr1:225692756-225695652	2897	none	none	15	3
12:13	148154-152073	chr1:225688773-225692692	3920	none	none	14	3
13:14	152153-154739	chr1:225686107-225688693	2587	none	none	412	242
14:15	154798-155331	chr1:225685515-225686048	534	none	none	309	215
1:2	459-85729	chr1:225755117-225840387	85271	none	none	360	91
2:3	85896-98060	chr1:225742786-225754950	12165	none	none	1432	536
3:4	98239-122505	chr1:225718341-225742607	24267	none	none	1083	344
4:5	122591-133578	chr1:225707268-225718255	10988	none	none	632	189
5:6	133947-135837	chr1:225705009-225706899	1891	none	none	277	146
6:7	135949-138243	chr1:225702603-225704897	2295	none	none	518	117
7:8	138549-140127	chr1:225700719-225702297	1579	none	none	262	201
8:9	140274-140429	chr1:225700417-225700572	156	none	none	661	245
9:10	140488-141284	chr1:225699562-225700358	797	none	none	410	260

Exon skipping junction read - raw count tables

Junction (A:B)	Gene coordinates	Chromosomal coordinates	Junction length	Donor (A)	Acceptor (B)	Counts Sknsh_LongpolyA_1	Counts Sknshra_LongpolyA_1
11:13	145194-152073	chr1:225688773-225695652	6880	none	none	614	263
5:7	133947-138243	chr1:225702603-225706899	4297	none	none	69	5

C



FIGURE 33 – **La visualisation RNA-Seq FasterDB. A.** Visualisation des données RNA-seq dans FasterDB. Cette visualisation se sépare en 3 blocs. Le premier représente le gène tel que défini dans FasterDB. Le deuxième représente les données RNA-Seq. Pour chaque échantillon, nous avons deux visualisations : la couverture des lectures sur le gène et les jonctions. La dernière partie permet de visualiser les différents transcrits du gène annotés dans FasterDB. **B.** Tableau de comptage des jonctions. Ce tableau est en deux parties. La première contient toutes les jonctions canoniques (entre deux exons génomiques consécutifs). La deuxième contient toutes les autres jonctions. C'est dans ce deuxième tableau que l'on peut trouver les jonctions excluant un ou plusieurs exons. Pour chaque jonction, le comptage de cette jonction dans chaque condition est affiché. Dans cet exemple, il y a deux conditions d'où deux colonnes de comptage. **C.** Légende pour la représentation des jonctions. Les jonctions sont représentées avec des couleurs différentes suivant la quantification de la jonction. Les jonctions supportées par très peu de lectures seront représentées en violet avec des traits très fin. Alors que les jonctions supportées par un très grand nombre de lectures seront représentées en rouge avec des traits très épais.

Cette visualisation qui est a destination des biologistes avec pas ou peu de connaissance en

informatique, est donc une alternative intéressante à IGV. En effet, avec cette visualisation, il n'est pas nécessaire de télécharger les données sur son propre ordinateur et ce dernier n'a pas besoin d'être puissant. Alors que pour utiliser IGV, il faut un ordinateur avec beaucoup de mémoire vive et suffisamment de place sur le disque dur pour pouvoir stocker les données.

E Validation de la méthode

Notre outil d'analyse de l'épissage a été validé sur des données générées par le laboratoire et sur des données de collaborateurs.

La première validation a été réalisée sur des données générées par mon équipe. Le RNA-Seq a été réalisé sur une lignée cellulaire de cancer du sein, MCF7, dans deux conditions. Dans la première condition, les deux ARN hélicases DDX5 et DDX17 ont été déplétées avec des siRNAs ("small interfering RNA"). Les siRNAs sont de petites séquences nucléotidiques qui vont cibler certains ARNs spécifiques grâce à la complémentarité de leur séquence. Les ARNs ciblés sont alors dégradés et l'expression du gène dont proviennent les ARNs est alors réduite. La deuxième condition est la condition contrôle. Le gène ciblé par le siRNA est le gène de la luciphérase de la drosophile (GL2) qui n'est donc pas présent dans les cellules humaines. Un seul séquençage a été réalisé pour chaque condition.

Des validations par RT-PCR ont été réalisées par Amandine Rey, doctorante dans l'équipe, sur des événements d'exon cassette trouvés par FaRLine. 38 événements d'épissage ont été sélectionnés avec des $\Delta\Psi$ variants de 60% à 10% en valeur absolue. Nous avons obtenu un très bon taux de validation avec 36 événements validés sur les 38 testés. Quelques résultats de PCR migré sur gel sont montrés dans la figure 34 B. A partir de ces gel, l'intensité des bandes représentant chacun des variants (inclusion et exclusion) ont été quantifiés avec le logiciel ImageJ. La quantification de chacune de ces bandes, nous permet ensuite de calculer un $\Delta\Psi$ en utilisant la formule suivants :

$$\Delta\Psi = \frac{inclusion}{inclusion + exclusion}$$

Nous avons ensuite comparé les $\Delta\Psi$ trouvé par les deux méthodes. Les résultats de quantification des événements par RNA-Seq sont similaires aux résultats trouvés par RT-PCR. On voit en effet que la corrélation des $\Delta\Psi$ du RNA-Seq et des RT-PCR est plutôt bonne avec un R^2 égal à 0.70 (Figure 34 A), alors que la RT-PCR n'est pas une méthode précise au niveau quantitatif.

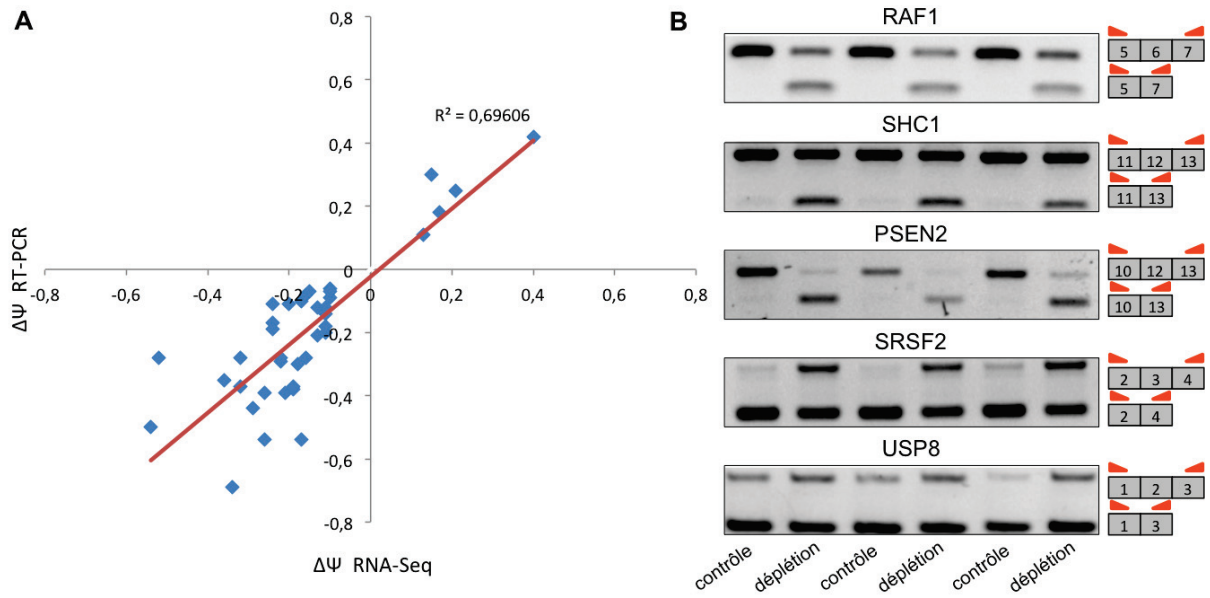


FIGURE 34 – **Validation expérimentale des événements de saut d'exon trouvés par FaRLine. A.** Corrélation des $\Delta\Psi$ calculés dans le RNA-Seq avec les $\Delta\Psi$ calculés en validation par RT-PCR. **B.** Gels de validation par RT-PCR de 5 exons cassettes régulés par la déplétion de DDX5 et DDX17. Toutes les validations ont été réalisées en triplicat.

La deuxième série de validations a été réalisées par nos collaborateurs d'I-Stem. Nous avons analysé pour eux un jeu de données RNA-Seq portant sur l'étude de l'effet d'une molécule thérapeutique, la biguanide metformine, sur le transcriptome. Des précurseurs cellulaires dérivés de cellules souches de patients atteints de DM1 ont été utilisés comme modèle cellulaire. Le plan expérimental comportait trois conditions : le contrôle et deux doses différentes de traitement à la metformine (10mmol/L et 25 mmol/L). Chaque condition a été séquencée en triplicat biologique. Les 20 événements de saut d'exon les plus fortement régulés avec la plus grande dose de metformine ont été testés par RT-PCR. Parmi ces 20 événements, 19 furent validés. Ces validations se trouvent dans la publication de *Molecular Therapy - Nucleic Acids* en annexe (dans la section I).

F Valorisation de la méthode

En plus des analyses qui ont permis de valider FaRLine, j'ai réalisé de nombreuses autres analyses avec ce pipeline.

Pour l'équipe, j'ai analysé des données publiques provenant du projet ENCODE. Notamment, FaRLine a permis de comparer l'épissage alternatif dans différents types cellulaires. La comparaison des lignée cellulaires de type épithéliale (cellules polarisées, organisées en couche dense et peu motiles) versus des lignées cellulaires de type fibroblastique (cellules fusiformes ou étoilées et

motiles) a été utilisé dans le manuscrit mis en annexe (partie II). J'ai également réalisé un grand nombre d'analyse sur des données générées par l'équipe. La plupart de ces données proviennent de lignées MCF7 et SH-SY-5Y (lignée de neuroblastome) avec diverses déplétions. Notamment, la déplétion des ARNs hélicases DDX5 et DDX17 qui sont un sujet important d'étude dans l'équipe a été étudié dans ces deux modèles cellulaires.

J'ai également réalisé de nombreuses analyses pour des collaborateurs. Notamment dans le cadre de la collaboration avec I-Stem, j'ai utilisé FaRLine pour analyser des données RNA-Seq sur des précurseurs cellulaires dérivés de cellules souches de patients atteints de DM1 traités ou non avec une autre molécule que la metformine.

J'ai réalisé plusieurs analyses de données de patients atteints de leucémie myélomonocytaire chronique (LMMC) dans le cadre d'une collaboration avec Eric Solary de l'Institut Gustave Roussy à Paris. J'ai comparé l'épissage alternatif chez des patients présentant ou non une mutation dans le gène SRSF2. J'ai également analysé des données RNA-Seq afin d'étudier l'épissage chez des patients atteints de LMMC qui répondaient plus ou moins bien à un traitement (composé d'agents déméthylants de l'ADN). Les patients répondant au traitement ont été comparés avant et après traitement à des patients ne répondant pas au traitement.

Une collaboration avec Nicolas Charlet Berguerand de l'IGBMC de Strasbourg m'a permis d'analyser l'épissage alternatif chez des patients atteints de DM1 dans 2 tissus différents (le coeur et le muscle squelettique).

Deux collaborations m'ont permis d'étudier des données de résistance à des thérapies anticancéreuses. La collaboration avec Martin Dutertre de l'Institut Curie à Paris, m'a permis d'analyser l'épissage alternatif dans une lignée cellulaire MCF7 rendue résistante à la Doxorubicine (molécule utilisée dans certaines chimiothérapies). De plus, dans la lignée rendue résistante, certains facteurs ayant été trouvés comme pouvant rétablir la sensibilité au traitement ont été déplétés pour pouvoir analyser leur effet sur l'épissage alternatif.

Le deuxième projet sur la résistance au thérapie a été réalisé en collaboration avec Béatrice Eymin de l'IAB de Grenoble. Le plan expérimental du RNA-Seq contenait deux lignées cellulaires de cancer du poumon (HCC827 et PC9). Chacune de ces deux lignées a été rendue résistante à des thérapies ciblées (gefitinib et dacomitinib). Chaque lignée rendue résistante a été comparée à la lignée cellulaire sensible d'origine.

La dernière collaboration a été réalisée avec Edouard Bertrand de l'IGMM de Montpellier. Di-

vers facteurs impliqués dans la biogenèse des ARNs ont été déplétés et j'ai comparé les données avec et sans déplétion. Cela a permis d'étudier le rôle de ces différents facteurs dans l'épissage alternatif.

Au cours de ma thèse, j'ai donc réalisé des analyses sur un vaste panel de données comprenant à la fois des lignées cellulaires et des cellules de patients, des données avec ou sans traitement en collaboration avec de nombreuses équipes différentes.

G Discussion

Au cours de ma thèse, une grande partie de mon travail fut de développer un pipeline d'analyse de données RNA-Seq adapté aux besoins de l'équipe. Pour cela, j'ai développé une nouvelle méthode d'analyse de l'épissage alternatif : FaRLine. Puis, j'ai mis en place l'automatisation du pipeline pour permettre de réaliser facilement et rapidement des analyses de données RNA-Seq, aussi bien au niveau de l'expression des gènes que de l'épissage. Les analyses tournent sur le cluster de calcul de l'école normale supérieure de Lyon, le PSMN. Ces outils sont régulièrement utilisés par différents bioinformaticiens de l'équipe pour lancer des analyses sur des données publiques, générées par l'équipe ou par des collaborateurs.
Les résultats de FaRLine ont été validés expérimentalement par l'équipe et par des collaborateurs. Les taux de validation sont particulièrement élevés (supérieurs à 90%). De plus, cette méthode a déjà été utilisée pour analyser un grand nombre de jeux de données.

FaRLine nécessite encore des améliorations. Notamment, il serait intéressant d'ajouter l'analyse de nouveaux types d'événements d'épissage, comme les premiers et les derniers exons alternatifs et les rétentions d'intron. L'analyse de ces événements permettrait d'être plus exhaustif pour l'étude de l'épissage alternatif. Il serait également intéressant de donner la possibilité à l'utilisateur de prendre en compte les lectures exoniques en plus des lectures jonctions pour la quantification. Cela permettrait très certainement de gagner en puissance dans l'analyse statistique.
De plus, un certain nombre de développements techniques sont encore nécessaire pour permettre la distribution de notre outil à la communauté scientifique. Notamment, il faudrait rendre l'installation de FaRLine sur un nouveau système plus facile.

II Comparaison d'une méthode basée sur l'alignement à une basée sur l'assemblage

A Publication soumise à Genome Research

Comme expliqué dans la section III.C de l'introduction, il existe de nombreux outils pour analyser l'épissage alternatif dans les données RNA-Seq. Pour réaliser une analyse, il faut choisir entre des méthodes basées sur l'alignement et d'autres basées sur l'assemblage.

La plupart du temps un utilisateur va utiliser une seule méthode ou lorsque quelqu'un utilise différentes méthodes, il va prendre uniquement les événements communs aux différentes méthodes pour la suite de ses analyses. Mais les événements trouvés uniquement par une méthode doivent-ils vraiment être ignorés ou ne sont ils pas manqués par une autre approche à cause de limites méthodologiques ? C'est la question que nous nous sommes posés dans notre étude.

Nous avons comparé FaRLine qui est basé sur l'alignement à une approche basée sur l'assemblage, nommée KisSplice. Ces deux méthodes permettent d'étudier localement les événements d'épissage et la comparaison a été réalisée uniquement sur les événements d'exon cassette. Pour réaliser la comparaison, nous avons utilisé un jeu de données public issue du projet ENCODE : une lignée cellulaire de neuroblastome, SK-N-SH, différentiée ou non par traitement à l'acide rétinoïque.

Une grande proportion des événements trouvés par les deux méthodes sont communs. En effet, parmi les sauts d'exon avec une expression minimum, 70% sont annotés par les deux approches. Mais nous avons noté des différences non négligeables. Notamment, l'approche basée sur l'alignement est plus sensible et permet donc de trouver des variants d'épissage avec une faible expression. De plus, elle prédit mieux les événements d'épissage pour des exons chevauchant des éléments répétés comme des éléments ALU. D'un autre côté, la méthode basée sur l'assemblage qui ne dépend pas des annotations, permet de prédire des événements d'épissage non annotés (nouveaux exons ou nouveaux sites d'épissage). Cette approche permet aussi de prédire des événements d'épissage dans les familles de gènes paralogues.

Ces événements ne peuvent pas être ignorés car ils sont validés expérimentalement. Et certains d'entre eux sont différentiellement régulés entre les deux conditions comparées.

Enfin, nous avons comparé nos outils avec d'autres méthodes publiées, comme Trinity, Cufflinks et MISO. Ces comparaisons nous ont permis de confirmer que les méthodes locales comme FaRLine et KisSplice étaient plus sensibles que les méthodes globales telles que Trinity ou Cufflinks.

Cela nous a également permis de confirmer avec d'autres méthodes les différences que nous avons observé entre une approche basée sur l'alignement et une autre basée sur l'assemblage.
Dans ce projet, je me suis occupée de l'analyse des données RNA-Seq avec FaRLine et MISO. J'ai également réalisé toutes les comparaisons entre les différentes approches, sélectionné les cas à valider expérimentalement et supervisé leur validation.

Toutes les références citées dans cet article sont présentes à la fin de l'article ainsi que les figures principales et supplémentaires.

Annotation and differential analysis of alternative splicing using *de novo* assembly of 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], Cyril Bourgeois[1], Didier Auboeuf[1,*], 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, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, 69622, Villeurbanne, France. Vincent.lacroix@univ-lyon1.fr

Didier Auboeuf, Laboratory of Biology and Modelling of the Cell, ENS de Lyon, 69007, Lyon, France. Didier.auboeuf@inserm.fr

Keywords = "*de novo* assembly", "alternative splicing", "transcriptome"

Abstract

Genome-wide analyses reveal 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). These two approaches are event-based, as they focus on the regions of the transcripts that vary in their exon content. We applied these methods to an RNAseq dataset from a neuroblastoma SK-N-SH cell line (ENCODE) differentiated or not using retinoic acid. We 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 families of paralog genes. The mapping-first approach allowed to find more lowly expressed splicing variants, and was better in predicting exons overlapping repeated elements. This work demonstrates that annotating AS with a single approach leads to missing a large number of candidates. We further show that these candidates cannot be neglected, since many of them are differentially regulated across conditions, and can be validated experimentally. We therefore advocate for the combine use of both mapping-first and assembly-first approaches for the annotation and differential analysis of AS from RNAseq data.

1 Introduction

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 [Pan et al., 2008, Wang et al., 2008]. The depth at which we can sample transcriptomes with next generation sequencing techniques opens the possibility not only to annotate splicing variants in physiological conditions, but also to detect which transcripts are differentially spliced across conditions.

This growing interest in splicing both as a fundamental process and because of its implication in pathologies [Scotti and Swanson, 2016, Edery et al., 2011, David and Manley, 2010] has been accompanied by an increasing number of methods aiming at analyzing RNAseq datasets [Trapnell et al., 2012, Wang et al., 2010, Robertson et al., 2010]. 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 [Steijger et al., 2013]. 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) [Reyes et al., 2013, Katz et al., 2010, Shen et al., 2012, Sacomoto et al., 2012].

Whether they are exon-centric or isoform-centric, methods to study splicing from RNAseq data can further be divided in two main categories [Martin and Wang, 2011]. The mapping-first approaches first map the reads to the reference genome and 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 data however still remains rare, although some studies have already shown its potential to detect novel splicing variants which play a central role in the studied disease [Dargahi et al., 2014, Freyermuth et al., 2016].

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 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 the same publicly available RNAseq dataset.

As a first step, we chose to compare pipelines that we developed in parallel in two teams, 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 benchmarked our methods to other classically used pipelines and were able to confirm the generality of our findings.

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.

From a general point of view, the combination of approaches we propose will enable researchers to extend significantly their list of candidates.

2 Results

2.1 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 EnsEMBL r75 annotations. In the mapping-first approach, reads are aligned to the reference genome using TopHat2. Mapped reads are then analyzed by FaRLine, in the light of the EnsEMBL r75 annotations.

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 where only junction reads are considered. 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 further developed and adapted jointly these two pipelines in order to minimize the discrepancies that could unnecessarily complicate our comparison.

2.2 The majority of frequent isoforms are found by both approaches

Applying KisSplice and FaRLine to the same RNAseq dataset (SK-N-SH cell lines treated or not with retinoic acid) generated by the ENCODE consortium, 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 (Figure 2 A). 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 full 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 found that for ASEs reported only by FaRLine, the number of reads supporting the minor isoform is much lower than in the other categories (Figure 2 B).

In order to further compare the mapping and assembly-first approaches, we decided to filter out candidates for which the minor isoform was sup-

ported by less than 5 reads or whose relative abundance was lower than 10% compared to the major isoform.

As expected, the proportion of candidates reported by both methods increased significantly. Approximately 70% of predicted skipped exons were now found by both approaches. (Figure 2C).

Furthermore, the estimation of their inclusion levels were very 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) (Figure 2D), the failure of one approach to detect them is likely not due to a lack of coverage.

Moreover, events from each of these 3 categories were validated by RT-PCR (Figure 3 and Supplementary Figure S1).

For all these cases, we patiently dissected the reasons why they could have been missed out by one approach. This enabled us to define 4 main categories (Figure 3A).

2.3 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 unannotated exons. The unannotated exon can be the skipped exon or one of its flanking exons. It can also be a subpart of a larger annotated exon, and hence be overseen (see Methods).

The reason why the mapping-first approach does not detect these events is twofold. First, the mapper may map the reads to an incorrect location, as junction discovery using short reads is a challenging task. This occurred

in 17% of the 1436 cases. Second, in the case where the mapper succeeds (83% of the cases), FaRLine failed to report the event because it relies on annotations. Among these 1199 cases, we distinguished 3 sub-categories of errors due to the annotation. Either the exon is unannotated (28%), one of its flanking exon is unannotated (8%) or both exons are annotated but no transcript combining them was annotated (45%). 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.

The downstream annotation of the events is then permissive, in the sense that annotations are used as an evidence, not as a guide. Alternative splicing events involving novel splice sites are clearly identified as such, and can be individually tested and experimentally validated. HIRA gene contain a novel exon, whose inclusion is supported by at least 20 reads on each junction (Figure 3B). This case was overseen by the mapping-first approach, FaRLine. The panel A of the supplementary figure S2 shows an example of an ASE not reported by FaRLine because the inclusion was not present in the transcripts.

The second category of splicing events identified by only one approach corresponds to paralog genes. Untangling the relation between alternative splicing and gene duplication is a difficult topic, subject to debate [Kopelman et al., 2005, Roux and Robinson-Rechavi, 2011]. 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 ma-

jority 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 paralog genes. 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.

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 635 cases belonging to this category, a large fraction (40%) 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 [Batzer and Deininger, 2002]. Most of these copies are located in introns and a number of them have been exonised [Lev-Maor et al., 2003, Sorek et al., 2004]. The reason why the mapping-first approach is able to identify these cases is because even though the read 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 S3). The assembly-based ap-

proach 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 (online supplementary figure `http://kissplice.prabi.fr/sknsh/graph_RAB5C_distance_3.html`). 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 S4). Increasing this threshold would lead to an increase in the sensitivity at the expense of the running time.

The fourth category of splicing events identified by only one approach corresponds to the 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 (Figure 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 Methods).

2.4 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 inclusion level of the exon 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 retinoic acid. For each detected event, we tested with KISSDE [Lopez-Maestre et al., 2016], 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 Figure 4, we found again that 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 (Figure 4C), with a DeltaPSI of 27%. The estimation of the DeltaPSI is overall very similar across the two approaches (Figure 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 S5 gives an example.

When compared to the splicing event annotated as reported in Figure 2, we noticed that the proportion of condition-specific events detected by only one method increased, for two main reasons. First, some ASE identified by both approaches were found as differentially included only by one method. This is again due to differences in the quantification of the inclusion levels, 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 hap-

pened to be condition specific. This is the case of an exon in NINL intron 5 (Figure 4D), only found by KisSplice because it was not annotated. This is also the case of SAR1B exon 3 (Figure 4E), only found by FaRLine because it overlaps with an Alu element.

The observation that many of the exons detected only by one method are differentially included across conditions confirms that these exons should not be discarded from the analysis. Focusing only on exons predicted by one approach may lead to miss splicing events which are central in the response to treatment.

2.5 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 dataset to check how its predictions overlapped with ours. As shown in Figure 5, 72% of predictions made by MISO were common to both FaRLine and KisSplice, 23% were only common with FaRLine, 2% were only common to KisSplice and the remaining 3% were specific to MISO. 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.

Beside 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 tran-

scripts.

The most widely used mapping-first and isoform-centric approach is Cufflinks [Trapnell et al., 2012] that we compared to FaRLine using the same dataset. As shown in Figure 5D, 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[Grabherr et al., 2011]. As shown in Figure 5B, 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 for each gene, it therefore largely under-estimates alternative splicing, especially inclusion/exclusion of short sequences.

2.6 Discussion

De novo assembly is usually applied to non-model species where no (good) reference genome is available. We show here that its usage, even when the annotated reference genome is available, offers a number of advantages. We name 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* about which exons should be found.

Using this strategy, we discovered many novel alternatively skipped exons, which were not found by traditional read mapping approaches (Figure 3). 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 and that a substantial number of new exons will be discovered. An example is the case of unannotated skipped exons

which overlaps 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 from paralogous genes (Figure 3). This is because mapping approaches discard reads mapping 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 [Poursani et al., 2016].

Conversely, we showed that some ASE were detected only by the mapping-first approach. As shown in Figure 2, 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 [Bacher and Kendziorski, 2016]. Therefore, computational analysis should not systematically discard lowly-expressed splicing variants and filtering these events should depend on the biological questions to address.

We also observed that the mapping-first approach better detects exons corresponding to annotated-repeat elements (Figure 3). 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 [Sorek et al., 2004, Shen et al., 2011].

When two methods have non-overlapping predictions, the temptation could be to focus on exons found by both approaches and discard the others. We argue that this would be a mistake, because these cases can be validated experimentally, and many of them correspond to regulated events, where the inclusion isoform is significantly up or down regulated in presence of a treatment.

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 of us. 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 by both approaches. The categories of challenging cases that we defined in Figure 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 had been assumed to play a minor role in transcriptomes. A number of recent work however argues in the opposite direction.

For exonised ALUs, paralog genes and genes with complex splicing, the possibility to sequence longer reads with third generation techniques [Tilgner

et al., 2014, Bolisetty et al., 2015] 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 splicing.

3 Material and Methods

3.1 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 introduced in [Sacomoto et al., 2012], kissDE was introduced in [Lopez-Maestre et al., 2016]. KisSplice2RefGenome and FaRLine are methods we introduce in this paper.

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

3.1.1 KisSplice

KisSplice is a local transcriptome assembler. As most short reads transcriptome assemblers [Grabherr et al., 2011, Schulz et al., 2012, Robertson et al., 2010], 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 cor-

respond to specific patterns in the DBG called bubbles. KisSplice's main algorithmic step therefore consists in enumerating all the bubbles in the graph built from the reads. The sequences corresponding to the two paths of each bubble are then aligned to the reference genome using STAR, and the result of the alignment is analysed using KisSplice2RefGenome to annotate the event.

3.1.2 Alternative splicing events are bubbles in the DBG

Supplementary figure S6 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.

3.1.3 Dealing with sequencing errors

As originally described in [Pevzner et al., 2004] and later in [Zerbino and Birney, 2008], 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.

3.1.4 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 [Tilgner et al., 2012]). Each intron is covered by few reads, but if a repeat is present in many introns, then this repeat will obtain a high coverage and will correspond to very dense subgraphs in the De Bruijn graph built from the reads. The traversal of such subgraphs to

enumerate all the bubbles they contain is long and mostly fruitless. We showed in [Sacomoto et al., 2014] 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.

3.1.5 Annotating the events with KisSplice2RefGenome

Bubbles found by KisSplice are mapped to the reference genome using STAR, with its default settings, which means that in case of multi-mappings, STAR reports all equally best matches. The mapping results are then analysed by KisSplice2RefGenome. At this stage, 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 S7). For exon skippings, 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 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 and is reported separately.

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 S8). The cases of exact repeats corresponds to recent paralogs.

3.1.6 FaRLine

FasterDB EnsEMBL r75 annotation

FasterDB RNAseq Pipeline, FaRLine, use the FasterDB-based EnsEMBL r75 annotation database. FasterDB is a database containing all annotated human splicing variants [Mallinjoud et al., 2014].

The genomic exons are defined by projecting the transcript exons (Supplementary Figure S9). First, the transcript exons are grouped by position. Then each group of exons define a projected exon with the following rules:

- The start is the smallest start of the non-first-exon of the group.
- The end is the highest 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 transcrits is an intron retention, the projected genomic exon is defined as a combination of the two exons the intron retained. In supplementary figure S9, the exons 5 and 6 and the intron 5 are considered as one unique exon. As events included in an exon are overseen, 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 [Trapnell et al., 2012].

```
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 [Li et al., 2009].

```
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

We wrote custom perl scripts, based on the FasterDB-based EnsEMBL r75 annotation database. 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 retrieve 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 4bp on read extremities are removed from the reads as we considered it is not sufficient to identify cor-

rectly 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. However, as the read size is only 76bp, this should not happen often. 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 read 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. So to define the flanking exons, we use the information contained in the transcripts and in the reads. We list each junction skipping an exon 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. Panel B of supplementary figure S2 gives an example of an ASE reported by FaRLine but not by MISO because the inclusion isoform is not annotated in the transcripts.

Comparison with STAR

We also mapped the reads with STAR, ran FaRLine on this alignments and compared the predicted skipped exon 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.

3.1.7 Differential analysis

Both pipelines perform ASE detection and quantification. The last step of the pipelines is the differential analysis of the expression levels of the variants. This task is performed using the kissDE [Lopez-Maestre et al., 2016] R package, which takes as input a table of read counts as in Figure S10, 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:

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

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 [Benjamini and Hochberg, 1995].

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) calculated for a pair of variants:

$$PSI_{condition} = \frac{counts_{variant1}}{counts_{variant1} + counts_{variant2}}$$

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 then we calculate their mean.

Finally, we output the DeltaPSI:

$$DeltaPSI = PSI_{condition1} - PSI_{condition2}$$

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 extremely difficult to perform any experimental validation.

3.2 SKNSH dataset

We downloaded a total of 959M reads from `http://genome.crg.es/encode_RNA_dashboard/hg19/`. 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.

3.3 Computational requirements

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 10Go 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/sknsh/`.

3.4 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 48h using 6μM 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 ngμL of initial RNA. PCR reactions were performed using GoTaq polymerase (Promega).

4 Acknowledgments

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 Rhône-Alpes (A.R.) and post-doctoral fellowships from Fondation ARC (M.P.L).

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.

References

[Bacher and Kendziorski, 2016] Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1):1.

[Batzer and Deininger, 2002] Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3(5):370–379.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

[Bolisetty et al., 2015] Bolisetty, M. T., Rajadinakaran, G., and Graveley, B. R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome biology*, 16:204.

[Dargahi et al., 2014] Dargahi, D., Swayze, R. D., Yee, L., Bergqvist, P. J., Hedberg, B. J., Heravi-Moussavi, A., Dullaghan, E. M., Dercho, R., An, J., Babcook, J. S., et al. (2014). A pan-cancer analysis of alternative splicing events reveals novel tumor-associated splice variants of matriptase. *Cancer informatics*, 13:167.

[David and Manley, 2010] David, C. J. and Manley, J. L. (2010). Alternative pre-mrna splicing regulation in cancer: pathways and programs unhinged. *Genes & development*, 24(21):2343–2364.

[Edery et al., 2011] Edery, P., Marcaillou, C., Sahbatou, M., Labalme, A., Chastang, J., Touraine, R., Tubacher, E., Senni, F., Bober, M. B., Nampoothiri, S., et al. (2011). Association of tals developmental disorder with defect in minor splicing component u4atac snrna. *Science*, 332(6026):240–243.

[Freyermuth et al., 2016] Freyermuth, F., Rau, F., Kokunai, Y., Linke, T., Sellier, C., Nakamori, M., Kino, Y., Arandel, L., Jollet, A., Thibault,

C., et al. (2016). Splicing misregulation of scn5a contributes to cardiac-conduction delay and heart arrhythmia in myotonic dystrophy. *Nature communications*, 7.

[Grabherr et al., 2011] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology*, 29(7):644.

[Katz et al., 2010] Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–1015.

[Kopelman et al., 2005] Kopelman, N. M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet*, 37(6):588–589.

[Lev-Maor et al., 2003] Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3'splice-site selection in alu exons. *Science*, 300(5623):1288–1291.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

[Lopez-Maestre et al., 2016] Lopez-Maestre, H., Brinza, L., Marchet, C., Kielbassa, J., Bastien, S., Boutigny, M., Monnin, D., Filali, A. E., Carareto, C. M., Vieira, C., Picard, F., Kremer, N., Vavre, F., Sagot, M.-F., and Lacroix, V. (2016). Snp calling from rna-seq data without a

reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research*.

[Mallinjoud et al., 2014] Mallinjoud, P., Villemin, J.-P., Mortada, H., Espinoza, M. P., Desmet, F.-O., Samaan, S., Chautard, E., Tranchevent, L.-C., and Auboeuf, D. (2014). Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome research*, 24(3):511–521.

[Martin and Wang, 2011] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682.

[Pan et al., 2008] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415.

[Pevzner et al., 2004] Pevzner, P. A., Tang, H., and Tesler, G. (2004). De novo repeat classification and fragment assembly. *Genome research*, 14(9):1786–1796.

[Poursani et al., 2016] Poursani, E. M., Soltani, B. M., and Mowla, S. J. (2016). Differential expression of oct4 pseudogenes in pluripotent and tumor cell lines. *Cell Journal (Yakhteh)*, 18(1):28.

[Reyes et al., 2013] Reyes, A., Anders, S., and Huber, W. (2013). Inferring differential exon usage in rna-seq data with the dexseq package.

[Robertson et al., 2010] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian,

J. Q., et al. (2010). De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912.

[Roux and Robinson-Rechavi, 2011] Roux, J. and Robinson-Rechavi, M. (2011). Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome research*, 21(3):357–363.

[Sacomoto et al., 2014] Sacomoto, G., Sinaimeri, B., Marchet, C., Miele, V., Sagot, M.-F., and Lacroix, V. (2014). Navigating in a sea of repeats in rna-seq without drowning. In *International Workshop on Algorithms in Bioinformatics*, pages 82–96. Springer.

[Sacomoto et al., 2012] Sacomoto, G. A. T., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., and Lacroix, V. (2012). KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC bioinformatics*, 13 Suppl 6(6):S5.

[Schulz et al., 2012] Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092.

[Scotti and Swanson, 2016] Scotti, M. M. and Swanson, M. S. (2016). Rna mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32.

[Shen et al., 2011] Shen, S., Lin, L., Cai, J. J., Jiang, P., Kenkel, E. J., Stroik, M. R., Sato, S., Davidson, B. L., and Xing, Y. (2011). Widespread establishment and regulatory impact of alu exons in human genes. *Proceedings of the National Academy of Sciences*, 108(7):2837–2842.

[Shen et al., 2012] Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.-x., Zhou, Q., Carstens, R. P., and Xing, Y. (2012). Mats: a bayesian

framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic acids research*, page gkr1291.

[Sorek et al., 2004] Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D., and Ast, G. (2004). Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Molecular cell*, 14(2):221–231.

[Steijger et al., 2013] Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., Bertone, P., Consortium, R., et al. (2013). Assessment of transcript reconstruction methods for rna-seq. *Nature methods*, 10(12):1177–1184.

[Tilgner et al., 2014] Tilgner, H., Grubert, F., Sharon, D., and Snyder, M. P. (2014). 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(27):9869–74.

[Tilgner et al., 2012] Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigó, R. (2012). Deep sequencing of subcellular rna fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas. *Genome research*, 22(9):1616–1625.

[Trapnell et al., 2012] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578.

[Wang et al., 2008] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B.

(2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.

[Wang et al., 2010] Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic acids research*, 38(18):e178–e178.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.

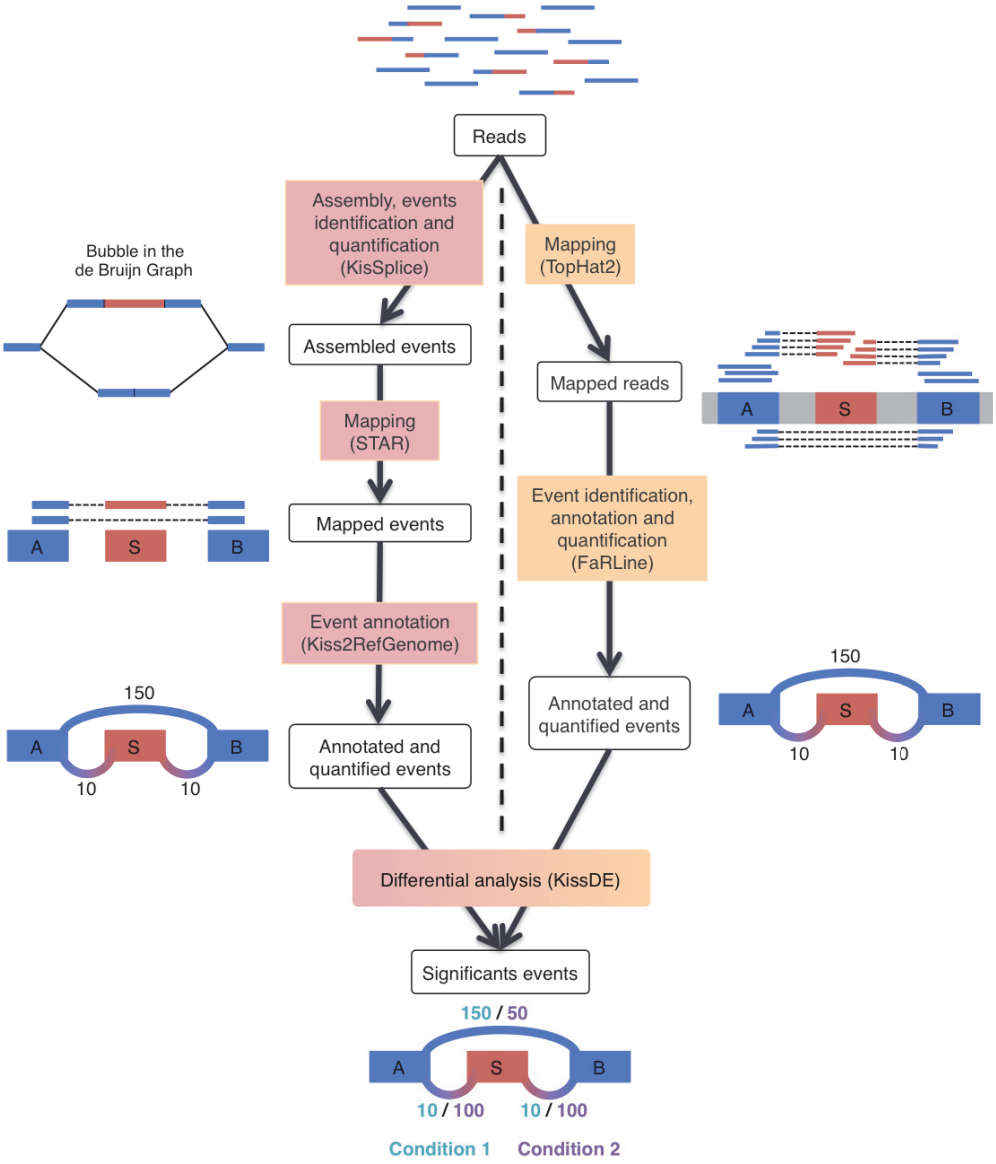


Figure 1: The two pipelines compared in this study: KisSplice and FaRLine. The first step of KisSplice is to assemble the reads and extract the splicing events. These events are then mapped back to the reference genome and classified by event type. The annotated and quantified events are then used for 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.

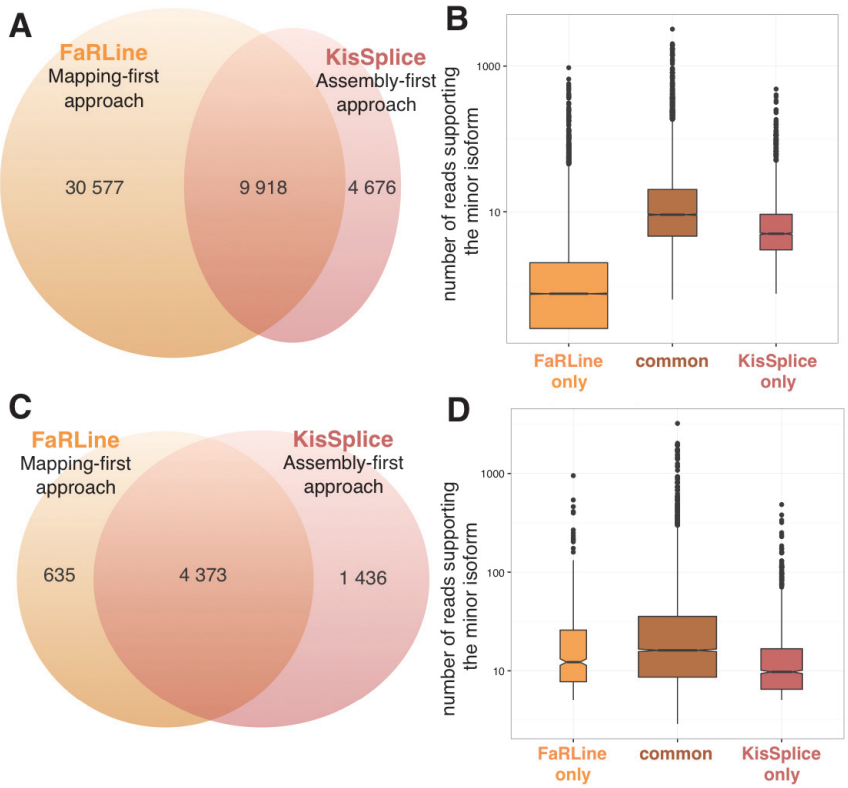


Figure 2: Comparison of the annotated ASE between assembly-first and mapping-first pipelines. A) Venn diagram of ASEs annotated by the two pipelines. FaRLine detected many more events than KisSplice. 68% of ASE annotated by KisSplice were also found by FaRLine and 24% of ASE annotated by FaRLine were 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 found only by FaRLine, ASE found by both pipelines and ASE found only by KisSplice. The number of reads supporting the minor isoform of the ASE found by FaRLine is globally much lower. C) Venn diagram of ASEs annotated by the two pipelines after filtering out the poorly expressed isoforms. The common events represent a larger proportion of the annotated events than previously: 87% of the ASE annotated by FaRLine and 75% 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 found only by FaRLine, ASE found by both pipelines and ASE found 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.

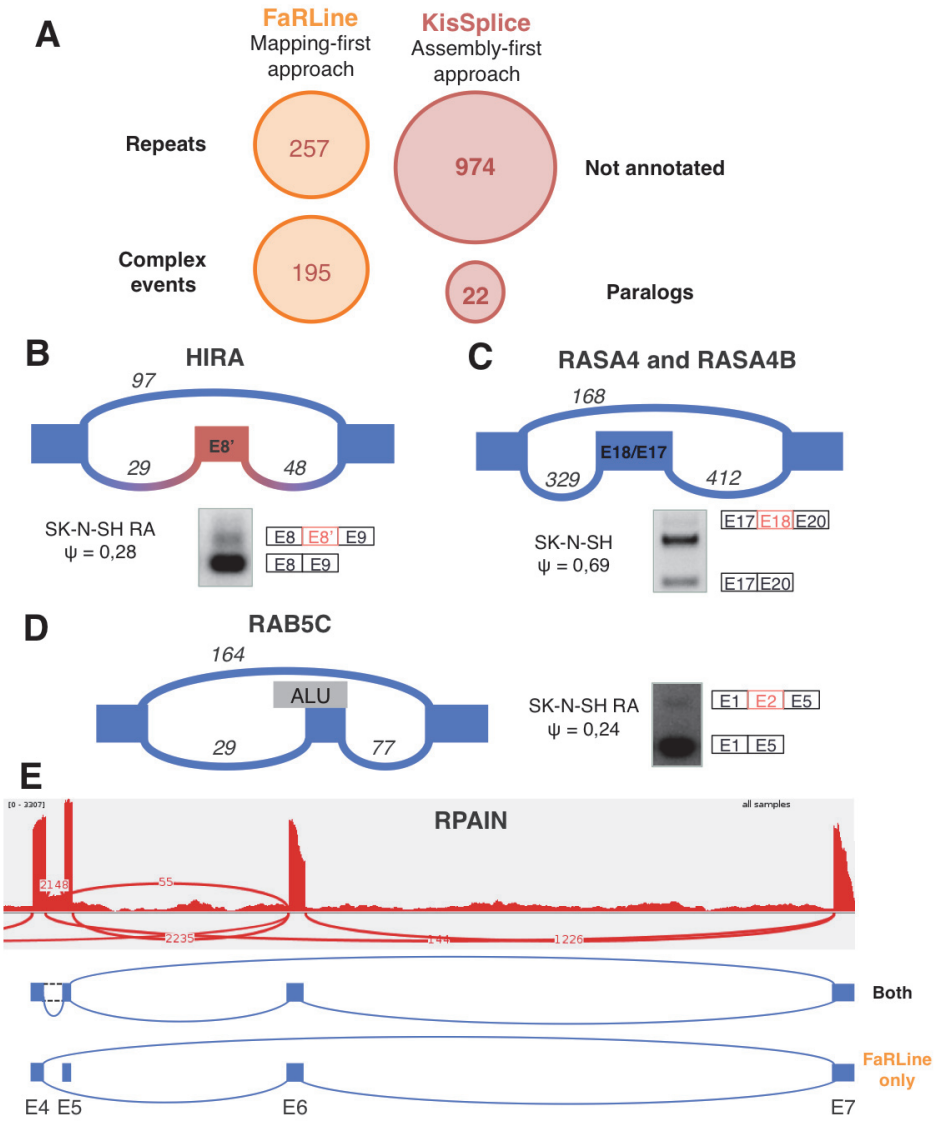


Figure 3: A) Categories identified explaining why some exons are detected by only one method. B) The new exon in intron 8 of the gene HIRA is an example of an exon not annotated in EnsEMBL r75. This event was found by KisSplice but not by FaRLine. C) RASA4 and RASA4B are 2 paralog genes. KisSplice detected 2 isoforms that could be produced by these 2 genes. FaRLine did not find 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. C) Exon 2 of the gene RAB5C is an example of exon skipping overlapping an Alu found only by FaRLine. The events in panel A to C were validated by RT-PCR. E) RPAIN contains a complex event with a lowly expressed isoform. This weakly expressed isoform was not found by KisSplice, while the other isoforms were found by both approaches.

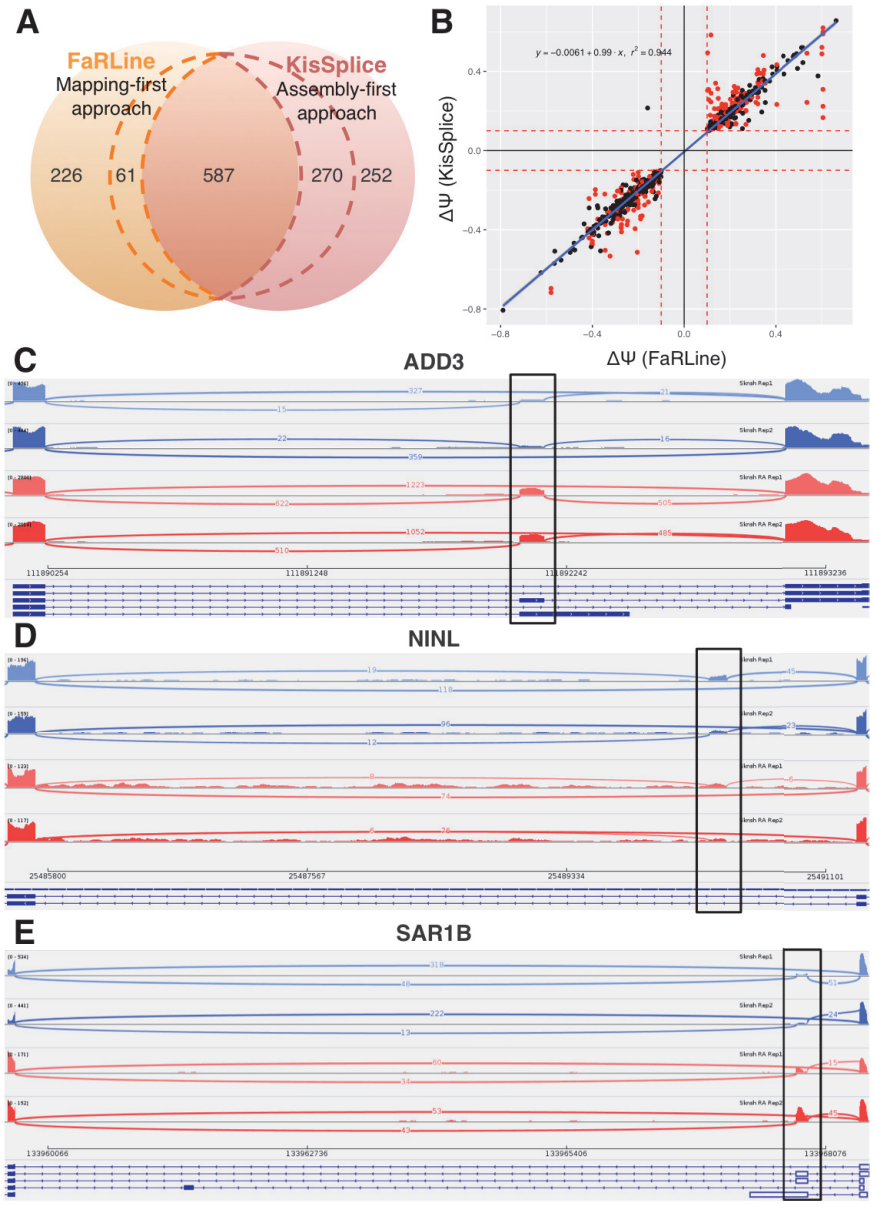


Figure 4: A) Condition-specific variants found by FaRLine, KisSplice or both methods. Within dashed lines are events identified by both approaches but detected as condition-specific by only one approach. B) DeltaPSI as estimated by KisSplice and FaRLine, for events found by both. The red points represent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

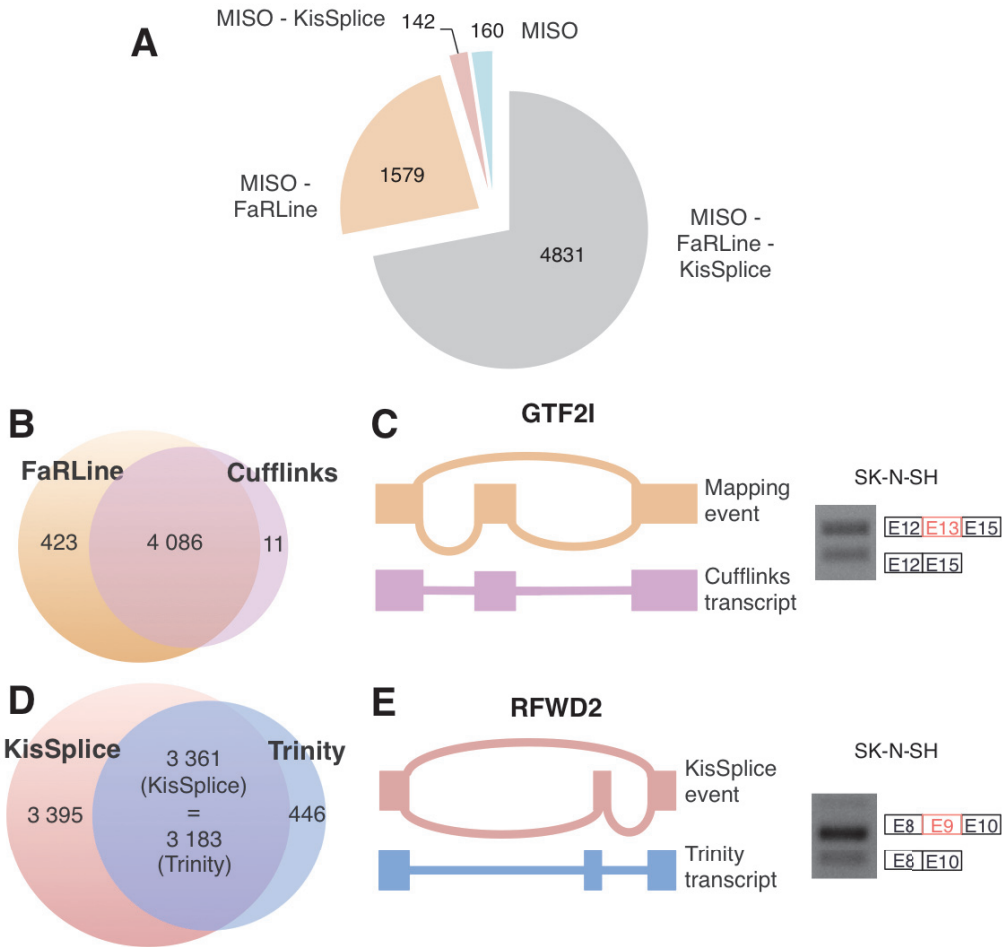


Figure 5: A) 72% of ASE found by MISO are also annotated by FaRLine and KisSplice. 23% of MISO's ASE are also annotated by FaRLine while only 3% of MISO's ASE are also annotated by KisSplice. Finally, only 2% of these ASEs are only annotated by MISO. B) Most of the events annotated by Cufflinks are found by FaRLine. C) GTF2I exon 13 is an example of an ASE annotated by FaRLine but not by Cufflinks. Indeed, Cufflinks only found the inclusion isoform. D) Most of the events annotated by Trinity are also found by KisSplice. But half of the ASE annotated by KisSplice are not found by the global assembler Trinity. E) KisSplice annotate an ASE in the gene RFWD2, while Trinity only found the inclusion variant. The events in panels C and E have been validated by RT-PCR.

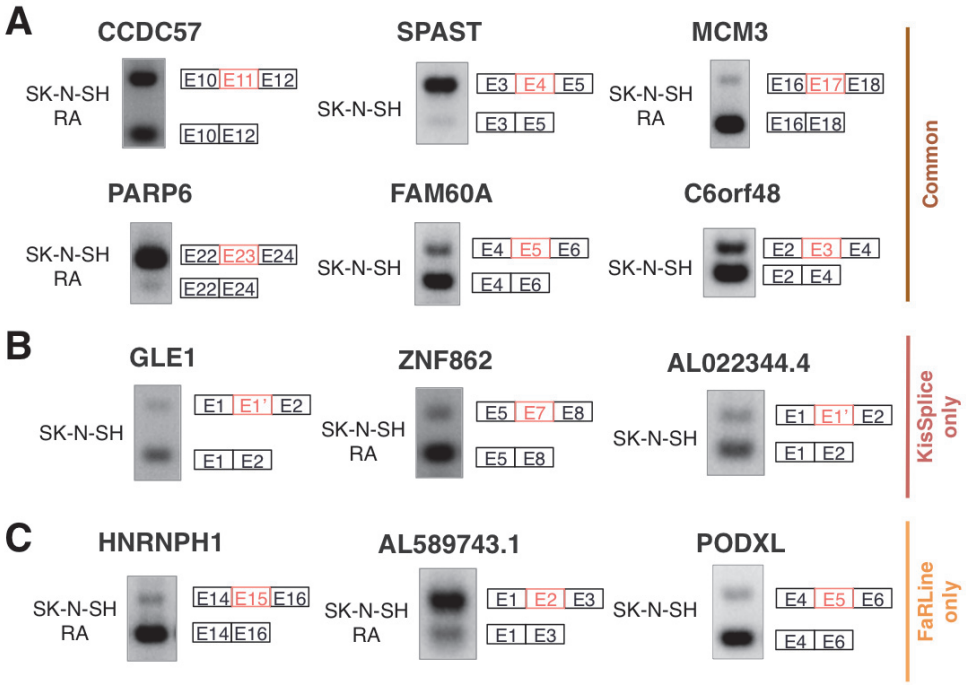


Figure S1: rt-PCR validations of events found by both approaches (A), only by KisSplice (B) and only by FaRLine (C).

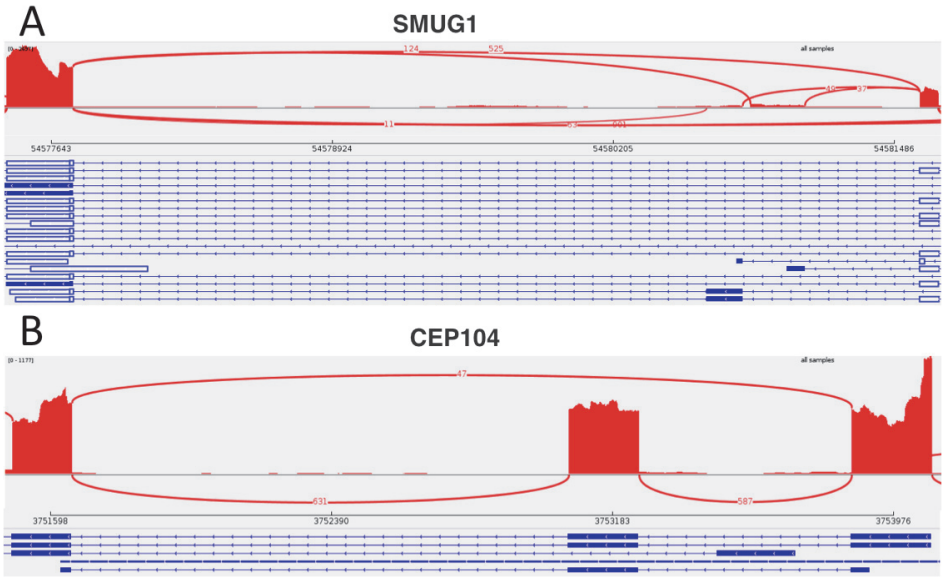


Figure S2: Examples of exon skipping inside a complex event. A) The exon 5 of SMUG1 gene is reported as skipped by KisSplice with exons 4 and 7 as flanking exons. This event is not found by FaRLine because the inclusion isoform is not annotated in the transcrits. B) Exon 12 of CEP104 gene is reported as skipped by FaRLine even if the exclusion isoform is not present in the annotation. However, MISO does not find this exon skipping.

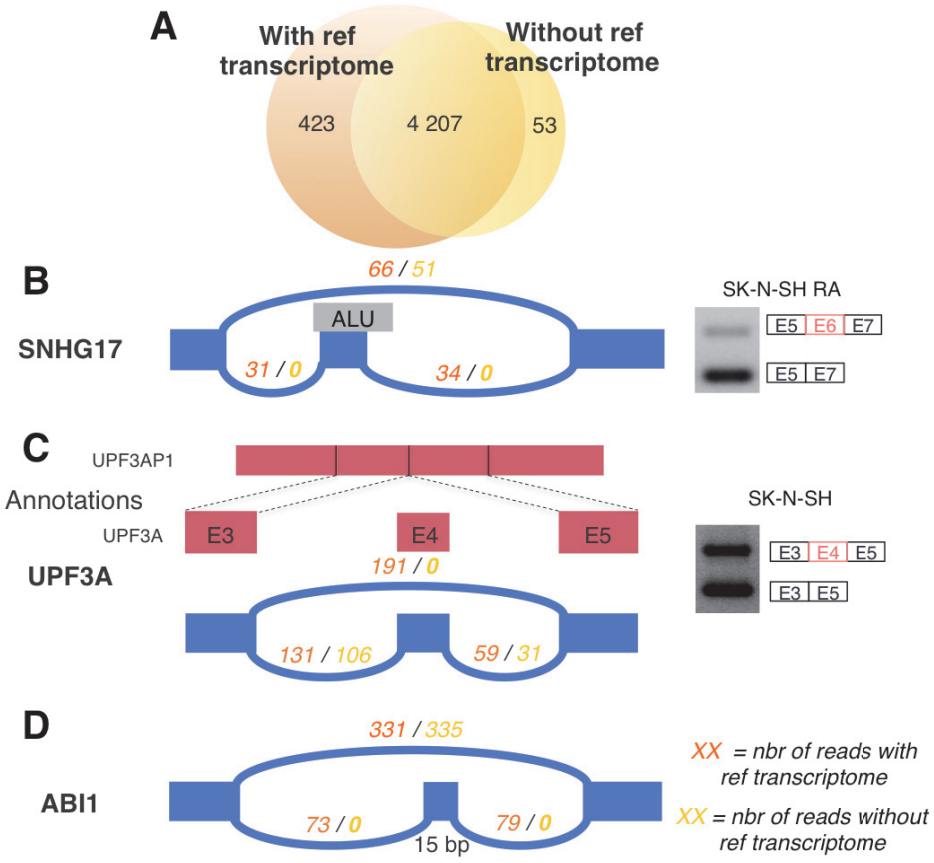


Figure S3: Comparison of the mapping-first approach FaRLine with or without an annotation provided to the mapper (i.e. with/without reference transcriptome). A) More ASE are annotated when an annotation available. Panels B to D show examples of events only found by the mapping-first method when an annotation is provided to the mapper. B) The first category, represented by the gene SNHG17, 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 20bp), like exon 5 of the gene ABI1, compose the third category.

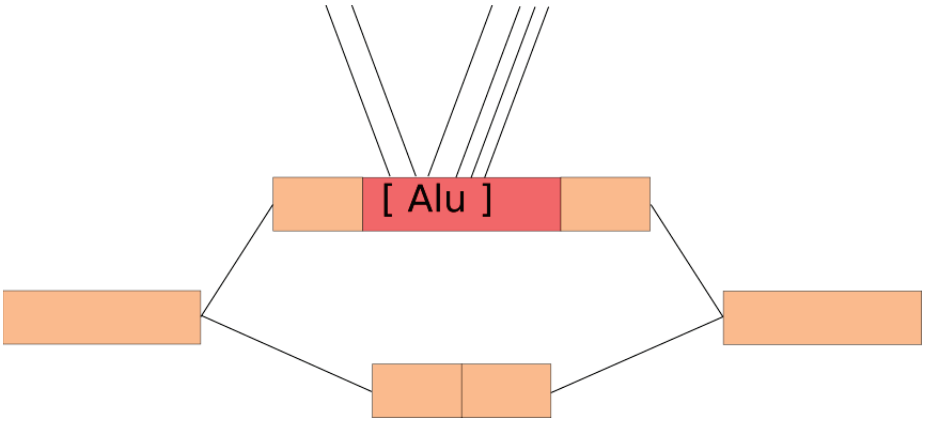


Figure S4: Example of a bubble containing an Alu. Repeated events such as Alu are expected to be present in several copies in the reads. Thus, when the graph is constructed, edges link different copies of Alu. Because a bubble with more than 5 edges within one of its paths is not enumerated by KisSplice, this case is not annotated by the assembly-first approach.

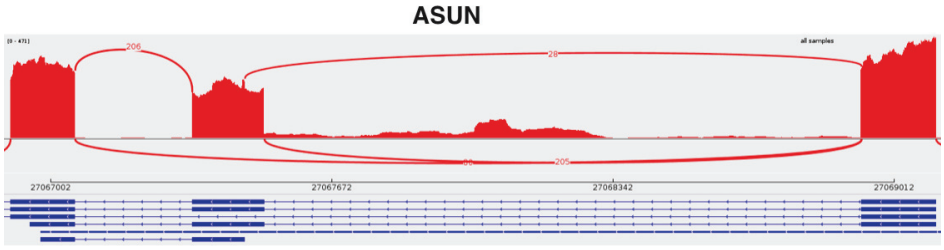


Figure S5: Example of an exon skipping with two alternative donor sites. It is reported as one event by FaRLine and two events by KisSplice.

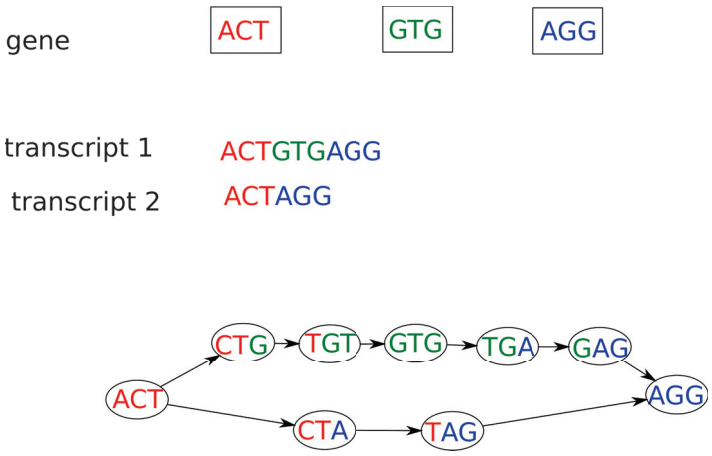


Figure S6: A schematic gene with three exons producing two alternative transcripts. The De Bruijn graph built from the sequences of the transcripts corresponds to a bubble. The upper path spells the skipped exon and its flanking junctions while the lower path spells the junction of the exclusion isoform and has a predictable length.

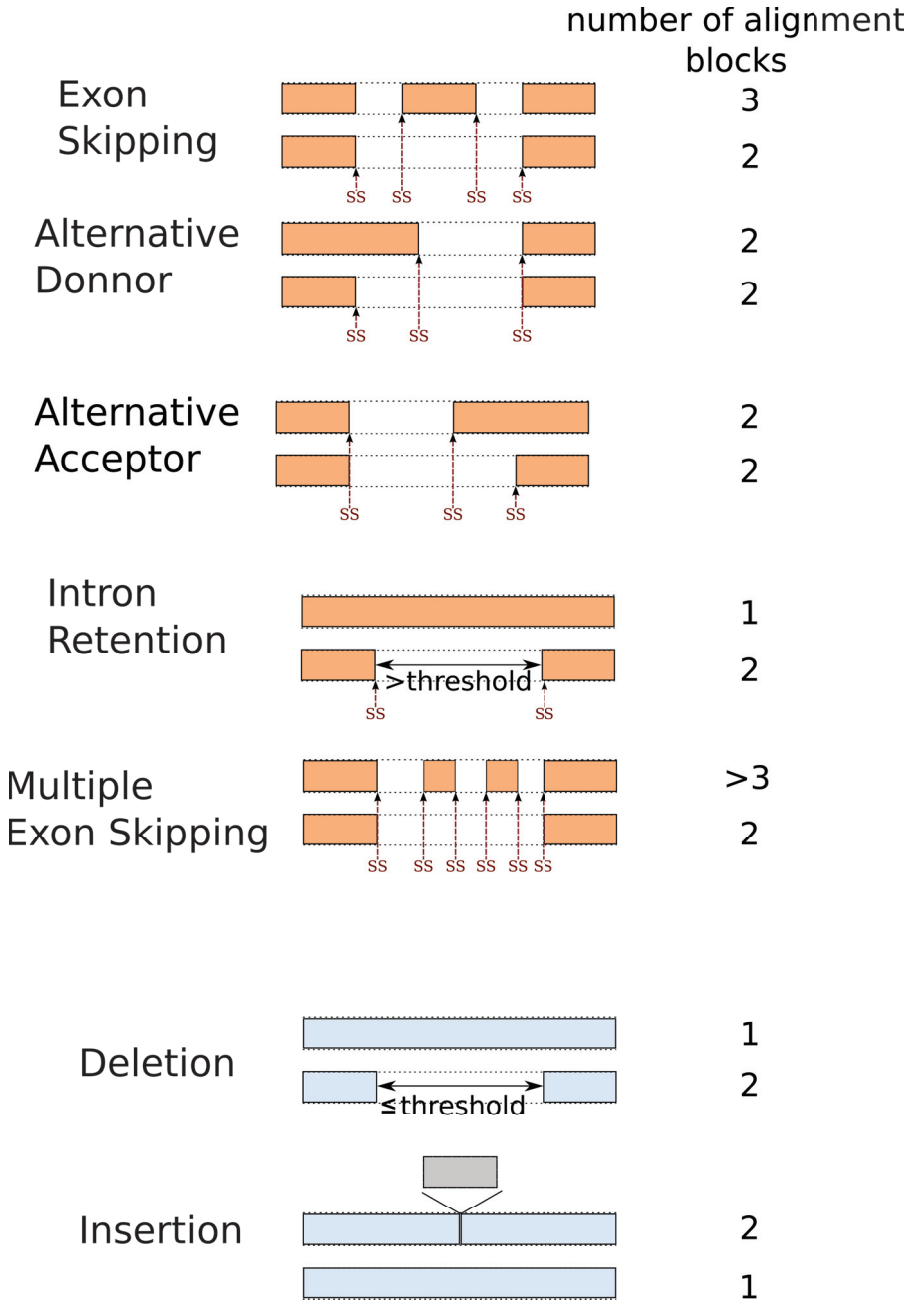


Figure S7: Classification of KisSplice events according to the number of blocks in which they map to the reference genome. Paths representing variants of an event are mapped on the reference. Spliced mapping results in blocks, events are then classified by KisSplice2RefGenome according to the block mapping patterns. (Putative) splice sites are noted by SS in red. In addition to alternative splicing, some indels are filtered through this step as they correspond to specific block patterns too.

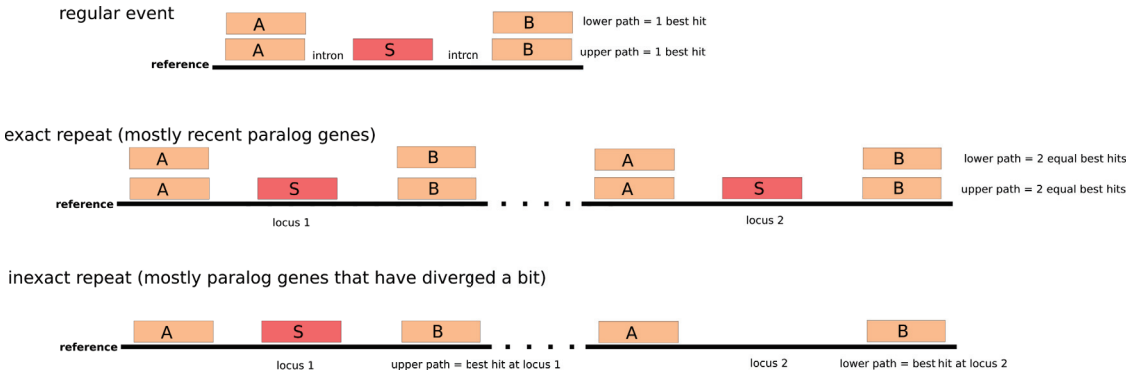


Figure S8: Dealing with repeats in KisSplice2RefGenome. If the two variants (i.e. paths) both map on different copies (exact repeat), we classify it as a recent paralog. On the contrary if each variant maps on a different locus, we consider the event as coming from an inexact repeat. This category represents mostly paralogs that have diverged.

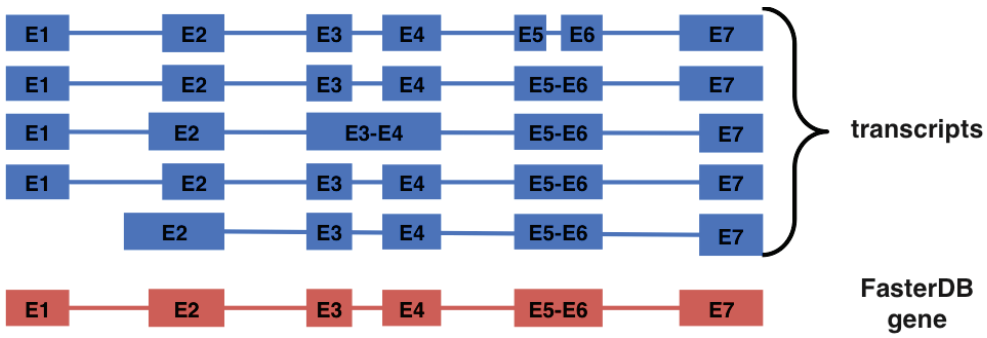


Figure S9: FasterDB exons are defined as the projection of the longer or most frequent exon in the transcripts (except for alternative first or last exons). The whole analysis done with FaRLine is based on these exons.

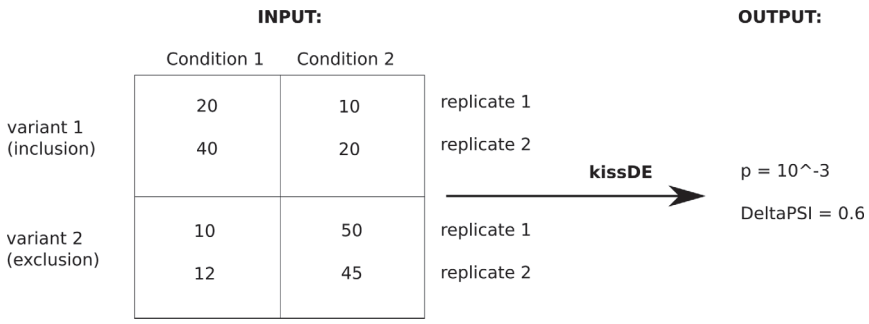


Figure S10: Input and output of the differential analysis. Counts for each replicate of each condition are computed by FaRLine or KisSplice. These counts together with the experimental plan are the input of kissDE. In the example, we show counts for one single event, in practice kissDE tests all events discovered by one method to spot the differential splicing events. Provided at least two replicates are available per condition, kissDE computes p-values and DeltaPSI per event, and results are ranked using these two metrics.

B Discussion

Cette comparaison montre que les méthodes basées sur l'assemblage et celles basées sur l'alignement sont complémentaires. En effet, malgré le grand nombre d'événements retrouvés par les deux approches, certaines catégories d'événements restent identifiables uniquement par une des deux méthodes.

Connaissant les catégories difficiles à trouver pour chacune de ces approches, on peut supposer que certaines catégories d'événement sont manqués par les deux approches. Par exemple, les exons ALU (ou contenant un autre type d'élément répété) non annotés ne sont sûrement trouvés par aucune de ces deux approches. D'où la nécessité d'améliorer encore les méthodes bioinformatiques ou de développer de nouvelles approches pour remédier aux limitations que nous avons mis en lumière dans cette étude.

D'autre part, cette comparaison nous a permis d'améliorer nos outils respectifs. En effet, au cours de la comparaison, nous avons pu trouver des événements trouvés par une méthode et qui aurait dû être trouvés par l'autre méthode également. Ces faux négatifs sont habituellement impossibles à connaître dans des données réelles. En les obtenant grâce à la seconde approche, nous avons pu modifier et perfectionner nos méthodes.

La prochaine étape est d'apporter une solution à ces limitations. La première solution serait de créer un pipeline alliant les deux approches en les faisant tourner en parallèle puis en fusionnant les résultats. Mais ce type d'outil serait très certainement gourmand en temps et en ressource de calcul. Une autre solution serait de tirer partie des avantages de chaque méthode. Pour cela, l'assemblage pourrait être utilisé pour enrichir les annotations qui seraient ensuite fournies à l'outil basé sur l'alignement des lectures. Mais cette solution ne permettrait pas de récupérer les événements d'épissage dans les familles de gènes paralogues. Une autre solution serait d'utiliser les annotations lors de l'assemblage. Cela permettrait sûrement de réduire le problème de sensibilité des approches basées sur l'assemblage des lectures, sans pour autant résoudre les limitations dues aux éléments répétés.

III Études de l'impact fonctionnelle de l'épissage alternatif

Une fois que l'on a identifié les variations d'épissage exprimés dans notre condition d'intérêt, l'analyse n'est pas terminée. Il faut encore comprendre les conséquences de ces variations sur le fonctionnement de la cellule. Pour cela, il faut trouver les conséquences de modification d'épissage au niveau de la protéine produite, mais aussi les répercussions sur les voies de signalisation.
Pour réaliser cet objectif, plusieurs outils ont été développés ou sont en cours de développement dans l'équipe. Ces outils sont principalement à destination des biologistes étudiant l'épissage alternatif.

A Annotation protéique à l'échelle des exons

Le premier outil appelé *Exon Ontologie* a été principalement développé par Léon-Charles Tranchevent, ancien post-doctorant dans l'équipe. Son objectif est d'annoter à l'échelle des exons les informations protéiques pour essayer de comprendre quelles pourraient être les conséquences fonctionnelles de modification d'épissage.
Pour cela, les exons génomiques de FasterDB ont été annotés avec les informations protéiques provenant de différentes ontologies et bases de données (Gene Ontology, Sequence Ontology, InterPro, ...). Les informations protéiques ont été organisées sous forme d'un arbre.
Les 8 grandes familles d'annotations à la base de l'ontologie sont : les domaines catalytiques, les domaines d'interactions, les domaines récepteurs, les domaines transporteurs, les motifs de localisations dans la cellule, les caractéristiques structurelles et les sites de modifications post-traductionnelles validés expérimentalement. Ces classes sont ensuite divisées en de nombreuses autres sous-catégories en suivant le fonctionnement d'une ontologie.
Une visualisation basée sur celle de FasterDB est associée à cet outil (`http://fasterdb.ens-lyon.fr/ExonOntology/`). Pour chaque gène on peut ainsi visualiser les différents domaines protéiques ou propriétés protéiques associés à chaque exons (Figure 35). Cette visualisation est simple et intuitive à utilisée. Il est également possible de fournir une liste d'exon pour récupérer les termes de l'exon ontologie qui leur sont associés. Enfin, il est aussi possible de calculer un enrichissement des différentes propriétés protéiques dans une liste d'exons.
Le manuscrit décrivant cet outil est disponible dans la partie II des annexes. Il a été soumis à *Genome Research* en juillet 2016.



FIGURE 35 – **Visualisation *Exon Ontologie*.** Cette visualisation est basée sur celle de FasterDB. On retrouve la représentation du gène dans le premier cadre (en haut de la page) et les différents transcrits annotés dans FasterDB dans le dernier cadre (en bas de la page). Entre les deux, on trouve des boutons permettant d'afficher les différentes grandes familles d'annotations d'*Exon Ontologie*. Ces annotations s'affichent avec une échelle différente que pour le gène et les transcrits. La taille des introns a été réduite afin de mieux visualiser les exons. En gris, on voit la représentation du gène avec cette nouvelle échelle et en dessous les différentes annotations. Ces dernières sont représentées en dessous de ou des exons les contenant. Dans cet exemple, on peut voir les différentes modifications post-traductionnelles (PTM) annotées dans le gène ENAH. On voit que lorsqu'une annotation chevauche plusieurs exons, elle est représentée en plusieurs blocs.

Exon Ontologie a été développé pour aider les biologistes lors de l'interprétation d'analyse de l'épissage alternatif dans des données haut débit (aussi bien puces à ADN que RNA-Seq). Grâce aux différentes possibilités qu'offre l'interface web, l'utilisateur peut décider d'étudier un seul événement d'épissage ou un ensemble d'événements. Il est intéressant d'analyser en même temps plusieurs événements d'épissage, car les conséquences fonctionnelles d'un seul événement d'épissage ne sont pas forcément suffisantes pour expliquer le phénotype cellulaire. En effet, si de nombreux événements ayant peu, voire pas d'effet sur le phénotype sont co-régulés et donc présents en même temps dans une cellule, les conséquences peuvent être très importantes sur le phénotype. C'est pour cela qu'il est important de replacer ces événements d'épissage dans leur contexte de voie de signalisation.

B Analyse des voies de signalisation

Le second outil permettant d'étudier l'impact de l'épissage alternatif sur la cellule est encore en cours de développement. Il permet d'intégrer dans les voies de signalisations provenant de wikipathways (`http://www.wikipathways.org/` [Pico et al., 2008]) les résultats d'analyse différentielle réalisés avec notre pipeline d'analyse : expression des gènes et épissage alternatif (uniquement les exons cassettes pour l'instant).

Cela est réalisé en colorant les gènes régulés. Le fond de la case représentant le gène est coloré, si le gène est régulé au niveau de son épissage (Figure 36 A). Si c'est une régulation au niveau de l'expression du gène, c'est la bordure de la case qui est colorée. La couleur indique le sens de régulation. En vert sont représentés les gènes sous-exprimés ou avec une exclusion de l'exon alternatif plus importante dans la condition testée par rapport à la condition contrôle. Et en rouge, les gènes sur-exprimés ou avec une inclusion de l'exon alternatif plus importante dans la condition testée par rapport à la condition contrôle. L'intensité de la couleur indique l'importance de l'effet : plus la couleur est pâle, plus l'effet est faible et inversement, plus la couleur est foncée, plus l'effet est fort. Pour chaque gène, il est possible qu'il y ait plusieurs événements d'épissage. Dans ce cas-là, le fond de la case est découpé en plusieurs rectangles de couleurs (un pour chaque événement d'épissage).

On utilise l'outil *Exon Ontologie* pour essayer d'évaluer l'importance des modifications d'épissage dans les voies de signalisations. Pour cela, chaque grande famille d'annotation d'*Exon Ontologie* a été associée à un symbole. Les symboles correspondants aux annotations présentes dans l'exon régulé sont affichés au coin du rectangle représentant le gène, comme on peut le visualiser sur la figure 36 A. Cette information permet de savoir d'un coup d'oeil s'il existe une annotation dans l'exon régulé. De plus, cela permet aussi de voir si un certain type de domaine semble particulièrement affecté. En parallèle, il est possible de réaliser une analyse fonctionnelle plus précise en utilisant *Exon Ontologie*.

Les fichiers représentants les voies de signalisation sont sous format gpml (Graphical Pathway Markup Language). Ce format est simplement un fichier de type xml avec des balises spécifiques. Ces fichiers peuvent être ouverts dans des logiciels libres comme PathVisio ou Cytoscape.

Cet outil est associé à une page intranet permettant aux membres de l'équipe de l'utiliser. Sur cette page, il est possible de choisir parmi des données RNA-Seq déjà analysées avec le pipeline de l'équipe ou de fournir de nouvelles données.

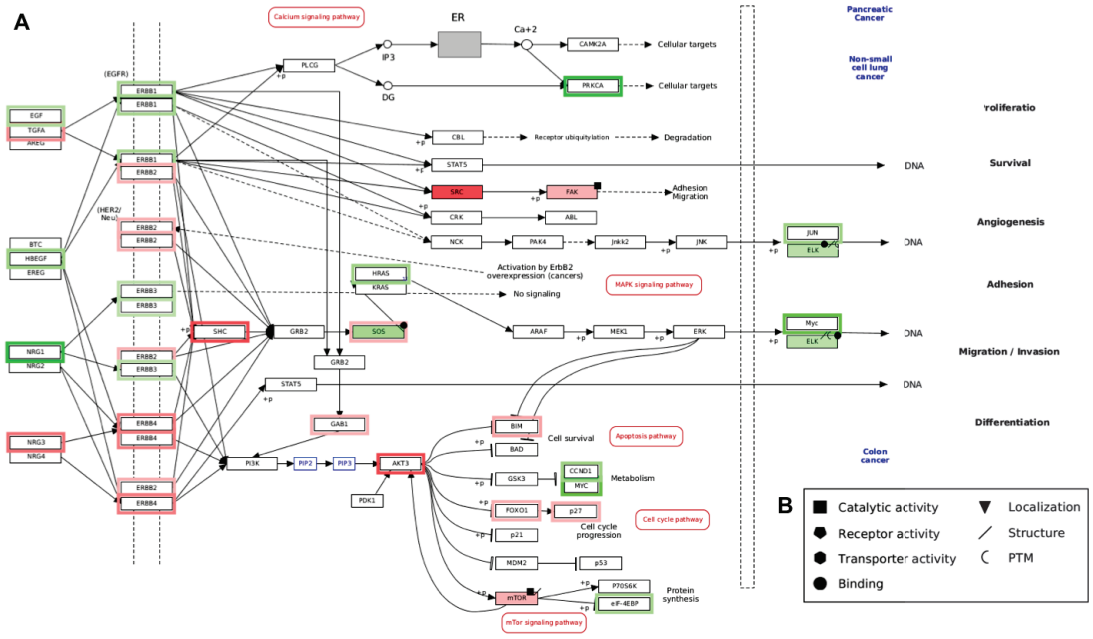


FIGURE 36 – **L'intégration des données RNA-Seq dans les voies de signalisation. A.** Exemple d'une voie de signalisation colorée avec les informations provenant de données RNA-Seq. La couleur de la bordure du cadre du gène indique le sens et l'intensité de la variation de l'expression du gène entre les deux conditions comparées. Alors que la couleur du fond du cadre du gène indique le sens et l'intensité de la variation de l'épissage du gène. Les symboles représentés sur le coin supérieur droit du cadre du gène indique les grandes catégories d'annotations présentes dans l'exon régulé. **B.** Correspondance des symboles avec les 8 grandes familles d'*Exon Ontologie*.

Comme indiqué au début de la section, cet outil est encore en développement. Il n'est accessible qu'aux membres de l'équipe. Il reste encore du travail pour le rendre diffusable à toute la communauté scientifique.

C Discussion

Les outils présentés ici permettent d'aller plus loin dans l'analyse de l'épissage alternatif à partir de données RNA-Seq. En effet, ils aident les biologistes à comprendre les conséquences de variations d'épissage. L'étude fonctionnelle des événements d'épissage se retrouve facilitée et elle est beaucoup moins orientée par les connaissances de la littérature. Cela peut potentiellement permettre de découvrir de nouvelles fonctions pour des gènes qui ont été peu étudiés jusqu'à aujourd'hui.

Ces outils permettent également une analyse plus générale. Les biologistes ne sont plus obligés de choisir des candidats dans la liste de résultats. Ils peuvent essayer de comprendre les conséquences de l'ensemble de ces variations d'épissage sur la cellule ou sur un processus biologique spécifique.

Chapitre 4

Conclusion et perspectives

L'épissage alternatif est un processus important et complexe. L'utilisation d'outils à la fois expérimentaux et bioinformatiques performants est donc capitale pour analyser de manière fine et la plus exhaustive possible ce processus. Le développement du séquençage haut-débit constitue une grande avancée pour l'analyse de l'épissage alternatif au niveau expérimental. Cette technologie permet aujourd'hui de générer une très grande quantité de données. Mais elle a fait surgir de nouveaux problèmes bioinformatiques.

Ces dernières années, de gros efforts de développement ont été faits pour créer des méthodes bioinformatiques permettant d'analyser le processus d'épissage. Mais il n'y a toujours pas de méthode consensus. De plus, la comparaison de méthodes que nous avons réalisé nous montre que pour l'instant les approches utilisées manquent certains types d'événements. Il reste donc des améliorations bioinformatiques à réaliser pour obtenir des analyses plus exhaustives lors de l'étude de ce mécanisme. Avec notre travail, nous avons mis en lumière les principaux points limitant de chaque approche, afin de permettre à la communauté bioinformatique de se pencher dessus.

Mon rôle au sein de l'équipe fut d'introduire et de mettre en place l'analyse des données RNA-Seq. Ce type d'analyse est maintenant utilisé en routine et va permettre à l'équipe de répondre à une variété de questions biologiques. FaRLine a été développé pour les biologistes afin de répondre à des problématiques biologiques. Il est donc important de revenir à ces problématiques initiales en remettant les résultats des analyses RNA-Seq dans le contexte des voies de signalisation.

Du point de vu biologique, l'épissage alternatif prend de plus en plus d'importance au sein de la médecine personnalisée. Cette approche thérapeutique qui se développe de plus en plus, notamment en cancérologie, consiste à traiter les patients de manière individualisée en fonction des spécificités génétiques et biologiques de leur maladie. Le mécanisme d'épissage s'insère dans cette approche thérapeutique par deux aspects.

L'épissage est un processus important de par son implication dans de nombreuses maladies. Ce mécanisme est de plus en plus considéré comme une bonne cible pour de nouvelles stratégies thérapeutiques. Une quantité croissante de molécules permettant de corriger les altérations de l'épissage est développée et testée. Certains traitements utilisent de petits oligonucléotides antisenses ciblant des séquences spécifiques d'ARN afin d'en modifier l'épissage [Southwell et al., 2012]. L'utilisation d'une séquence complémentaire au site d'épissage d'un exon va induire l'exclusion de cet exon. Alors que l'utilisation d'une séquence complémentaire à un élément "silencer" va

induire l'inclusion de l'exon. Ce type de thérapies a été testé pour de nombreuses maladies affectant l'épissage, comme l'ataxie télangiectasie [Du et al., 2007] ou la démence frontotemporal avec Parkinsonisme [Kalbfuss et al., 2001]. D'autres thérapies sont beaucoup moins ciblées et utilisent des molécules ayant un effet beaucoup plus global sur l'épissage. Notamment, un certain nombres de molécules anticancéreuses ont un effet inhibiteur ou modulateur sur l'épissage [Singh and Cooper, 2012]. De plus, il a été montré que certaines molécules actuellement commercialisées ont un effet sur l'épissage alternatif. Par exemple, la metformine, qui est utilisée pour traiter le diabète de type 2 module des événements d'épissage qui sont altérés chez les patients atteint de DM1 [Laustriat et al., 2015]. Le repositionnement thérapeutique de certaines molécules pourrait donc être une solution pour traiter certaines maladies impliquant l'épissage alternatif. Mais pour que ce type de thérapie fonctionne, il faut réussir à comprendre et contrôler le plus possible la régulation de l'épissage, afin de minimiser les effets non spécifiques qui pourraient entraîner des effets secondaires problématiques.

Aujourd'hui nous savons que des problèmes de résistance aux thérapies peuvent être dus à l'épissage alternatif. Nous avons des outils qui permettent d'inférer les conséquences de modifications d'épissage au niveau protéique et d'intégrer ces informations dans les voies de signalisation. La prochaine étape est d'utiliser la modélisation mathématique de voies de signalisation pour tenter de prédire les possibles résistances à des thérapies. De nombreux modèles de résistance à des thérapies ont été développés, notamment des résistances à des thérapies ciblées pour traiter différents types de cancers [Sahin et al., 2009, Sameen et al., 2016]. Mais ces modèles ne prennent en compte que l'expression des gènes et non pas l'épissage. Afin de mieux anticiper les possibles résistances à des traitements, il faut ajouter à ces modèles l'aspect qualitatif de l'expression des gènes : l'épissage alternatif.

Chapitre 5

Annexes

I Application à des données de collaborateurs

L'article ci-dessous a été publié en 2015 dans *Molecular Therapy - Nucleic Acids* avec nos collaborateurs de l'I-Stem (institut des cellules souches pour le traitement et l'étude des maladies monogéniques) à Evry.

L'étude portait sur le repositionnement thérapeutique de la biguanide metformine, une molécule actuellement utilisée dans l'hyperglycémie chez les individus atteints de diabète de type II. En effet, dans cet article, nous montrons que la metformine a un effet sur l'épissage alternatif. En particulier, elle permet de corriger certains défauts d'épissage observés dans des cellules fibroblastiques issues de patients atteints de DM1. Comme expliqué dans l'introduction (section I.C), la DM1 est une pathologie impliquant l'épissage alternatif et la correction de certains défauts d'épissage pourrait peut être permettre d'améliorer les conditions de vie des patients.
Pour étudier l'impact de cette molécule sur l'épissage alternatif, une analyse large échelle de RNA-Seq a été réalisée. Des précurseurs cellulaires dérivés de cellules souches de patients DM1 ont été utilisés comme modèle cellulaire. Le plan expérimental comportait trois conditions : le contrôle et deux doses différentes de traitement à la metformine (10mmol/L et 25 mmol/L). Le séquençage a été réalisé sur un séquenceur illumina avec une librairie *paired-end* (2x101pb). Chaque condition a été séquencé en triplicat biologique avec en moyenne 80 millions de lectures par échantillon.
Pour cette étude, j'ai réalisé l'analyse de l'épissage alternatif avec notre outil FaRLine et l'analyse de l'expression des gènes avec DESeq2. Les prédictions bioinformatiques ont été ensuite validées expérimentalement par les biologistes de l'I-Stem par RT-PCR dans les mêmes types de cellules qui ont servis à faire le séquençage RNA-Seq et dans des myoblastes (cellules souches responsables de la formation des muscles squelettiques) de patients DM1.

Citation: Molecular Therapy—Nucleic Acids (2015) 4, e262; doi:10.1038/mtna.2015.35
Official journal of the American Society of Gene & Cell Therapy All rights reserved 2162-2531/15

www.nature.com/mtna

In Vitro and *In Vivo* Modulation of Alternative Splicing by the Biguanide Metformin

Delphine Laustriat[1], Jacqueline Gide[1], Laetitia Barrault[1], Emilie Chautard[2,3], Clara Benoit[2], Didier Auboeuf[2], Anne Boland[4], Christophe Battail[4], François Artiguenave[4], Jean-François Deleuze[4], Paule Bénit[5,6], Pierre Rustin[5,6], Sylvia Franc[7], Guillaume Charpentier[7], Denis Furling[8], Guillaume Bassez[9], Xavier Nissan[1], Cécile Martinat[10], Marc Peschanski[10] and Sandrine Baghdoyan[10]

Major physiological changes are governed by alternative splicing of RNA, and its misregulation may lead to specific diseases. With the use of a genome-wide approach, we show here that this splicing step can be modified by medication and demonstrate the effects of the biguanide metformin, on alternative splicing. The mechanism of action involves AMPK activation and downregulation of the RBM3 RNA-binding protein. The effects of metformin treatment were tested on myotonic dystrophy type I (DM1), a multisystemic disease considered to be a spliceopathy. We show that this drug promotes a corrective effect on several splicing defects associated with DM1 in derivatives of human embryonic stem cells carrying the causal mutation of DM1 as well as in primary myoblasts derived from patients. The biological effects of metformin were shown to be compatible with typical therapeutic dosages in a clinical investigation involving diabetic patients. The drug appears to act as a modifier of alternative splicing of a subset of genes and may therefore have novel therapeutic potential for many more diseases besides those directly linked to defective alternative splicing.

Molecular Therapy—Nucleic Acids (2015) **4**, e262; doi:10.1038/mtna.2015.35; advance online publication 3 November 2015
Subject Category: Antisense oligonucleotides; Therapeutic proof-of-concept

Introduction

Alternative splicing of RNA is a key mechanism in increasing complexity of mRNA and protein metabolism. Imbalances in this splicing process may thus affect the progression of various human diseases. Identifying compounds capable of modulating this imbalance therefore provides new and interesting therapeutic perspectives.[1] Splicing is a conserved mechanism controlled by the spliceosome, a complex composed of five small nuclear RNAs (U1, U2, U4, U5, and U6) that assemble with proteins to form small nuclear ribonucleoproteins (snRNPs).[2] The production of alternatively spliced mRNAs is regulated by a system of trans-acting proteins that bind to cis-acting sites on the primary transcript itself. Each of these RNA-binding proteins has quite widespread effects on a number of genes. This sheds doubt on the ability of chemical compounds to target these factors in such a way as to have beneficial effects, without inducing concomitant deleterious consequences. One way to address this potential problem would be to focus on the repositioning of FDA-approved compounds that may affect alternative RNA splicing. It has already been demonstrated that marketed drugs such as clotrimazole, flunarizine, digitoxin, pentamidine, and manumycin A can also affect the alternative splicing machinery, raising the exciting prospect that the efficacy of disease-specific therapies may be enhanced by medications that target alternative splicing machinery.[3–6]

Within this framework, we were interested in the hypothesis that metformin, one of the most commonly prescribed antidiabetic drugs, downregulates the expression of a small subset of ribonucleic acid-binding proteins.[7] The effect of metformin on splicing machinery was explored in the pathological context of myotonic dystrophy type 1 (DM1), a model of spliceopathy where metformin is used to treat insulin resistance in DM1 patients.[1,8] DM1 is characterized by a defect in the alternative RNA splicing machinery, where 80% of the splicing alterations are related to the nuclear sequestration of the RNA-binding protein MBNL1 on myotonin protein kinase gene (*DMPK*) transcripts containing an abnormal expansion of CUG repeats in the 3′ UTR.[9–14] These nuclear aggregates also promote CELF1 hyperactivation,[15] and the concomitant deregulation of these two splicing factors promotes the alteration of alternative splicing in various genes that have been linked to symptoms of DM1.[16–18]

We therefore explored the consequences of metformin application with DM1-related abnormalities in alternative RNA splicing, using an *in vitro* model based on a human embryonic stem cell line (hESC) derived from an embryo characterized as a DM1-gene carrier during a preimplantation genetic diagnosis. This cell line was already instrumental in revealing alterations of the expression of several genes associated with functional disturbances in DM1.[19–21] Our results confirmed the ability of metformin to modify alternative splicing events, including some that are defective in DM1.

[1]CECS/AFM, Evry Cedex, France; [2]Centre de Recherche en Cancérologie de Lyon, INSERM U1052, Centre Léon Bérard, Lyon, France; [3]Université Lyon 1, CNRS, UMR 5558, INRIA Bamboo, Villeurbanne, France; [4]Centre National de Génotypage, Institut de Génomique, CEA, Evry, France; [5]INSERM UMR 1141, Hôpital Robert Debré, Paris, France; [6]Université Paris 7, Faculté de Médecine Denis Diderot, Paris, France; [7]Centre Hospitalier Sud Francilien and CERITD, Evry Cedex, France; [8]Sorbonne Universités, UPMC Université Paris 06, Centre de Recherche en Myologie, INSERM UMRS974, CNRS FRE3617, Institut de Myologie, Paris 75013, France; [9]GH Henri Mondor, Inserm U955, Université Paris Est, Créteil, France; [10]INSERM/UEVE UMR 861, Evry Cedex, France. Correspondence: Sandrine Baghdoyan, INSERM U861, I-Stem, Genopôle Campus 1, 5 Rue Henri Desbruères, 91030 Evry Cedex, France. E-mail: sbaghdoyan@istem.fr
Keywords: alternative splicing; AMPK; Metformin; myotonic dystrophy type 1; RBM3
Received 15 May 2015; accepted 22 September 2015; advance online publication 3 November 2015. doi:10.1038/mtna.2015.35

Results

Effects of metformin treatment on RNA-binding protein expression

The downregulation of transcripts encoding five ribonucleic acid-binding proteins (RBM3, SRSF1, SFPQ, SRSF6, and RBM45) by metformin in the millimolar range was identified by Larsson *et al.*[7] Concordant with this study, we confirmed that metformin induced a statistically significant decrease of RBM3 in DM1 and wild-type mesodermal precursor cells (MPCs) differentiated from DM1 and control hESCs at a dose of 25 mmol/l (**Figure 1** and **Supplementary Figure S1a**). In contrast, metformin treatment did not result in a statistically significant alteration of expression of SRSF1, SFPQ, SRSF6, and RBM45 (**Supplementary Figure S1b**) or MBNL1[22] and CELF1, both of which are involved in the pathological mechanisms of DM1 (**Figure 1**).

The effect of metformin treatment on cell viability, toxicity, apoptosis, and proliferation was monitored in DM1 MPCs exposed to a range of metformin doses. Treatments with the drug for 48 hours did not affect viability, cytotoxicity, or apoptosis up to doses of 35 mmol/l. Treatments for 24 hours with increasing doses of ionomycin and staurosporine, used as positive controls, induced as expected a rapid increase

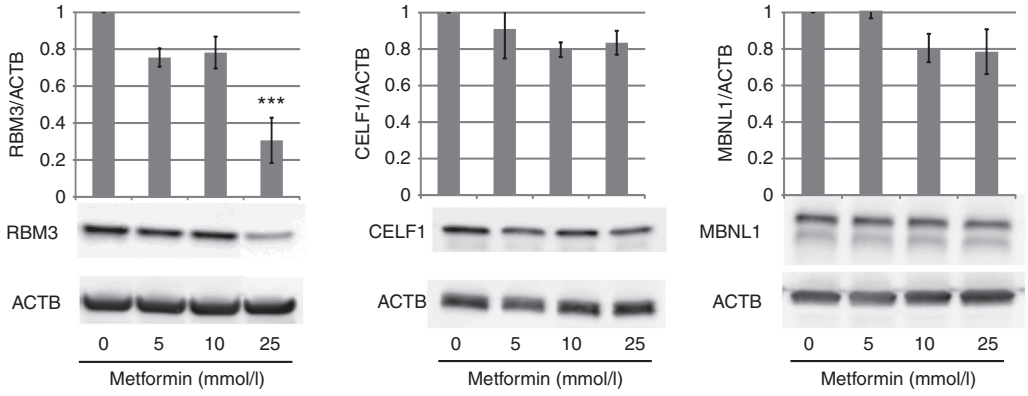


Figure 1 Metformin treatment modulates expression of the RBM3 RNA-binding protein. Western blot analysis of myotonic dystrophy type I mesodermal precursor cells treated for 48 hours with different doses of metformin demonstrated selective downregulation of RBM3, but not of CELF1 and MBNL1. Data (mean + SD) were analyzed with analysis of variance and the Steel-Dwass all pairs *post hoc* test. ***$P < 0.001$.

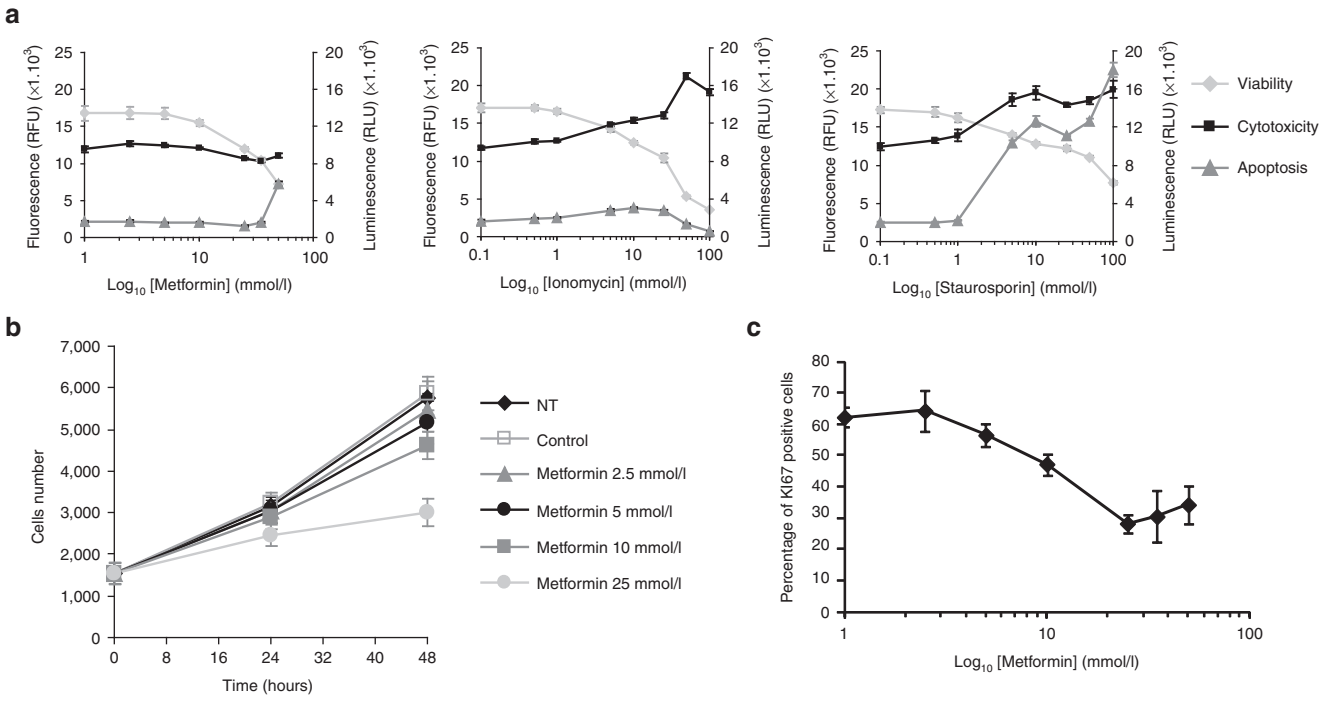


Figure 2 Effect of metformin treatment on myotonic dystrophy type I (DM1) mesodermal precursor cell (MPC) proliferation, apoptosis, and cytotoxicity. (**a**) DM1 MPCs were incubated with vehicle or increasing concentrations of metformin for 48 hours or staurosporine and ionomycin for 24 hours before measurement of viability, apoptosis, and toxicity with the ApoTox reagent. (**b,c**) DM1-mutated MPCs were treated with a range of metformin doses. Total cell numbers were determined at 0, 24, and 48 hours. Percentage of cells expressing the Ki-67 proliferation marker was analyzed after 48 hours of treatment. NT, not treated; RFU, relative fluorescence unit; RLU, relative luminescence unit.

of cytotoxicity and apoptosis, respectively[23,24] (**Figure 2a–c**). Proliferation analysis of DM1 MPCs showed that metformin treatment tended to promote a cytostatic effect at a dose of 10 mmol/l, with a greater effect at 25 mmol/l (**Figure 2b**). This was correlated with a progressive decrease in the number of cells expressing the Ki-67 proliferation marker (**Figure 2c**).

Effects of metformin on DM1-associated splicing defects

The consequences of metformin treatment were next analyzed on DM1-associated splicing defects in MPCs derived from DM1 hESCs. Changes in *INSR* exon 11 were studied using reverse transcription–PCR (RT–PCR), as the two isoforms of this gene were readily expressed in MPCs. Metformin corrected *INSR* exon 11 splicing defects in DM1 MPCs, increasing the inclusion from 18% (±2.3) to 34% (±1.8) at a dose of 25 mmol/l (**Figure 3a**). A similar shift towards an increased proportion of *INSR* exon 11 inclusion was also observed in wild-type MPCs, demonstrating the ability of metformin to regulate alternative splicing independently of the presence of the DM1 mutation. This analysis was extended to two other well-described alternative RNA splicing events associated with DM1, *TNNT2* exon 5 and *Clcn1* exon 7a.[17,25] This was based on the use of minigenes because the two isoforms of these genes are only expressed in heart and skeletal muscle.[25,26] The transient transfection of the minigenes in DM1 MPCs treated with 25 mmol/l metformin showed a statistical lowering of the percentage of inclusion of *TNNT2* exon 5 inclusion (from 51 ± 0.5% to 17 ± 0.4%) and *Clcn1* exon 7a inclusion (from 36% to 23 ± 2.03%) to levels similar to those quantified in wild-type MPCs (22 ± 1.5% and 20 ± 1.5%, respectively) (**Figure 3b**).

Overall effects of metformin treatment on alternative RNA splicing

To better understand the overall effects of metformin on gene expression and alternative RNA splicing, DM1 MPCs treated with 25 mmol/l metformin were analyzed using deep RNA sequencing. Treatment with 10 mmol/l of metformin was also included in this analysis as this dose induced a slight effect on *INSR* splicing (**Figure 3a**). A total of 63 and 1,171 genes in DM1 MPCs were regulated with an absolute log2-fold change of ≥1 ($P \leq 0.05$) in response to 10 and 25 mmol/l metformin, respectively. In these sets of genes, biological processes corresponding to cell cycle, response to DNA damage, cytoskeleton and ATP binding were enriched (**Table 1** and **Supplementary Tables S1** and **S2**). Downregulation of *RBM3* transcript was also detected in DM1 MPCs treated with 25 mmol/l metformin with fold changes of 0.33 (adjusted *P* value: 1.28×10^{-7}) confirming our previous observations (**Figure 1**). Analysis at the exonic level identified variations in 95 and 416 exons regulated above 10% ($P \leq 0.05$) at drug concentrations of 10 and 25 mmol/l, respectively (**Figure 4a**; **Supplementary Figure S2a** and **Supplementary Table S3**). Eighty-nine common splicing events were found to be deregulated at both concentrations. Of the 20 splicing events most highly regulated (>20%) by 25 mmol/l metformin, 19 were confirmed by RT–PCR in DM1 MPCs

Table 1 Biological process enriched in genes and splicings regulated by metformin

Biological process	Nb of genes	*P* value[a]	Fold enrichment
1,171 genes regulated by 25 mmol/l metformin (absolute log2-fold change > 1)			
Cell cycle	140	5.0 E-40	33.09
ATP binding	115	3.3 E-05	6.38
Cytoskeleton	112	1.1 E-05	8.51
Response to DNA damage stimulus	55	1.6 E-10	10.37
416 splicing events regulated by 25 mmol/l metformin (>10%)			
Cytoskeleton	53	3.3 E-04	4.26
Nuclear lumen	48	3.1 E-02	2.32
Kinase	30	1.6 E-02	1.76
RNA binding	24	3.2 E-02	1.87

[a]One tail Fischer exact probability value used for gene-enrichment analysis, using DAVID software and the human genome as reference.

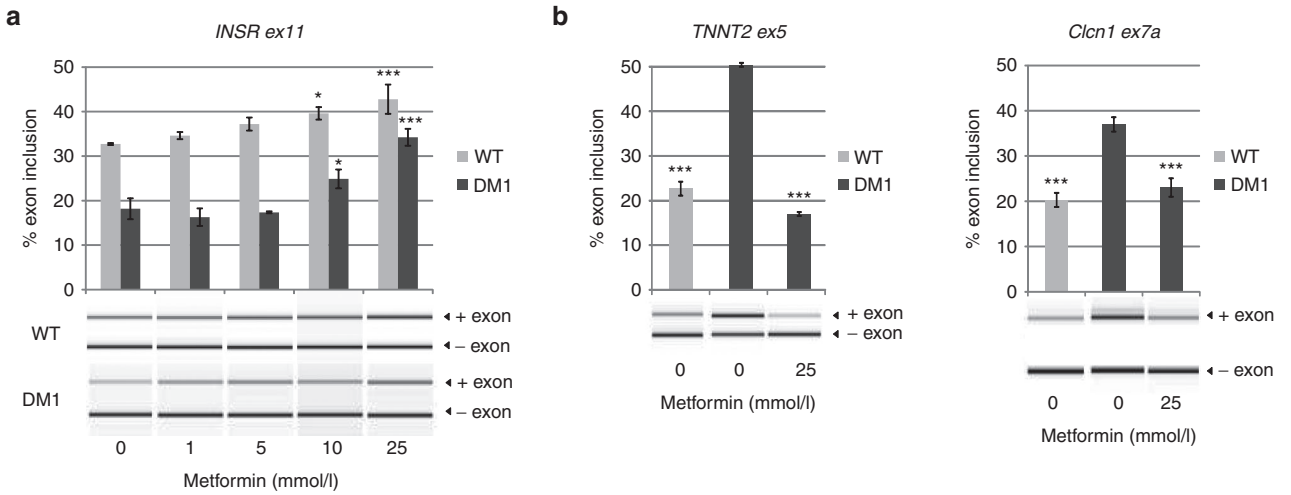


Figure 3 Characterization of metformin treatment on DM1-associated splicing defects in DM1-mutated mesodermal precursor cells (MPCs). (**a**) Alternative splicing of *INSR* exon 11 was analyzed with reverse transcription–PCR (RT–PCR) in wild type and DM1 MPCs treated with a range of metformin doses for 48 hours. (**b**) Alternative splicing of *TNNT2* exon 5 and *Clcn1* exon 7a were analyzed with RT–PCR in wild type and DM1 MPCs. Transfected DM1 MPCs were also treated for 48 hours with vehicle or 25 mmol/l metformin. Data (mean + SD) were analyzed with analysis of variance and the Kruskal–Wallis *post hoc* test. *$P < 0.05$, ***$P < 0.001$. DM1, myotonic dystrophy type I; WT, wild type.

a

Metformin	Splicing events modulated >10%
10 mmol/l	95
25 mmol/l	416
10 and 25 mmol/l	89

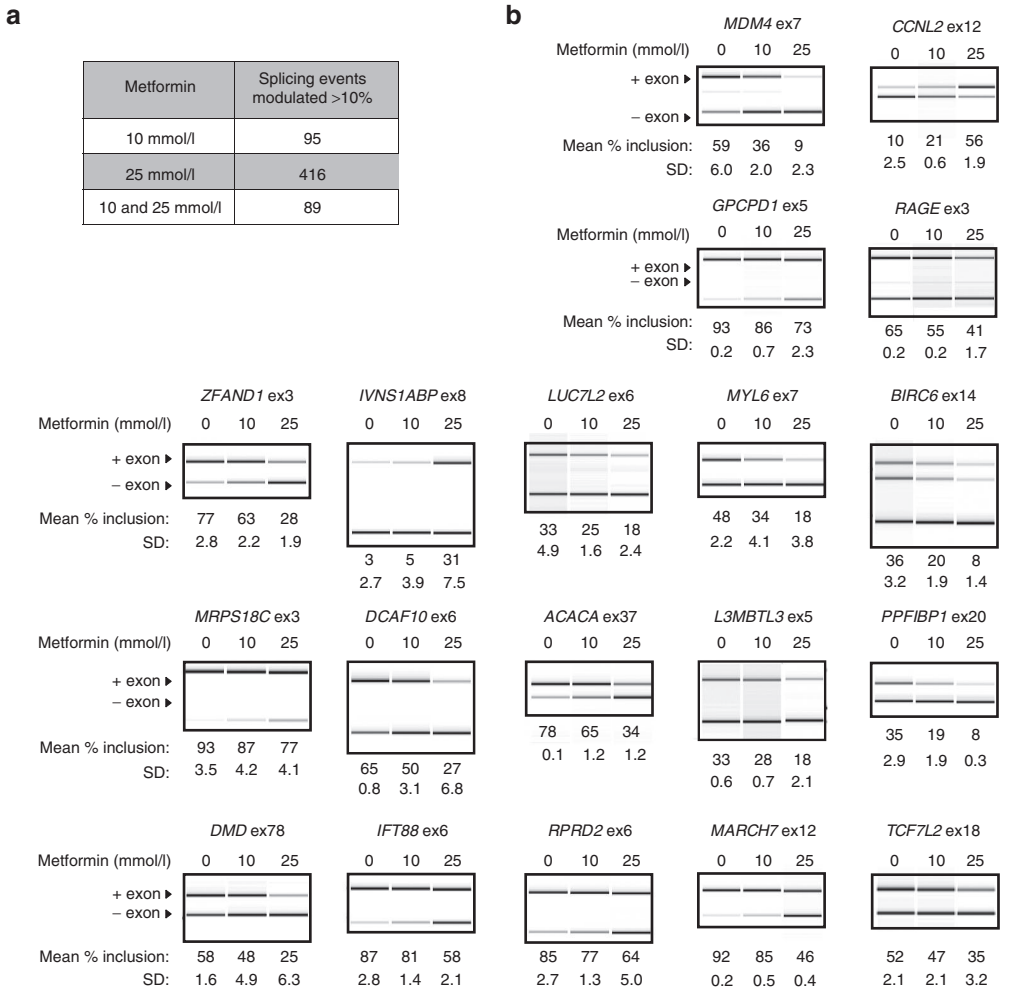


Figure 4 Genome-wide analysis of alternative splicing regulation by metformin. (**a**) Number of splice events modified by metformin treatment at 10 or 25 mmol/l. The number of common splice events modified at both concentrations is also indicated. (**b**) Reverse transcription–PCR (RT–PCR) validation of alternative splicing events identified by RNA sequencing in myotonic dystrophy type I (DM1) mesodermal precursor cells (MPCs) treated for 48 hours using 10 and 25 mmol/l of metformin. (**c**) Ten of the splicing events most regulated by metformin were analyzed by RT–PCR in DM1 MPCs transfected for 48 hours with a siRNA specific for *RBM3* or exposed for 48 hours to treatment with 25 mmol/l metformin. Data (mean + SD) were analyzed with analysis of variance and the Kruskal–Wallis *post hoc* test. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

(**Figure 4b**). Gene set enrichment analysis of the 416 splicing events regulated at 25 mmol/l identified sets of genes involved with the cytoskeleton, nuclear lumen, RNA binding, or with kinase activity (**Table 1**). Interestingly, most of the deregulated genes are also enriched in genes involved with the cytoskeleton (**Table 1**). None of the deregulated genes are associated with an alternative splicing modulation.

As our initial results indicated that metformin led to a reduced expression of RBM3, possible involvement of this splice factor in the regulation of alternative splicing by metformin was explored

using a siRNA transfection approach on 10 of the newly identified splicing events most altered by drug treatment. Thus, we compared the effects of metformin and transient extinction of *RBM3* in DM1 MPCs. Downregulation of *RBM3* expression induced similar results to metformin treatment for 7 out of 10 splicing events. These results strongly suggested that *RBM3* contributes to the mechanism of action of metformin on alternative splicing (**Figure 4c** and **Supplementary Figure S2b,c**).

As *RBM3* is described to be alternatively spliced, we also tested the possibility that the decreased expression of RBM3 induced by metformin was not associated to an effect of metformin on the expression of one of the alternate splice isoforms of *RBM3*. First, we analyzed, by RT–PCR, the expression profile of the 11 transcripts for *RBM3* that are described in ENSEMBL data bank in the presence and absence of metformin treatment. Among those, we detected six *RBM3* variant transcripts independently of metformin treatment (**Figure 5a**). We have next quantified the expression of each variant by quantitative PCR. Our results indicate that metformin treatment does not modify alternate splicing of *RBM3* transcripts (**Figure 5a**,**b**). In parallel, as two additional transcripts for *RBM3* have been described as candidates for nonsense-mediated mRNA decay process in NCBI data base, we also analyzed their expression after treatment with nonsense-mediated mRNA decay inhibitors, such as cycloheximide.[27] We first validated the effect of different concentrations of cycloheximide on the decay of unstable mRNA such as *c-FOS*. Our results indicate that, independently of the dose of cycloheximide used, the two *RBM3* transcripts are not expressed in our cell cultures in presence or not of metformin (**Supplementary Figure S3**). Altogether, our results indicate that the decreased *RBM3* expression observed after metformin treatment is not correlated to a change in alternative splicing of *RBM3* transcript.

Molecular mechanisms involved in the regulation of alternative RNA splicing by metformin

We next seek to understand the signaling pathways by which metformin treatment leads to a modification of splicing factor expression and consequently a modification of alternative splicing. The well-known antidiabetic action of metformin has been correlated to the inhibition of complex I of the respiratory chain,[28] raising the intracellular AMP/ATP ratio, a signal that finally triggers the downstream activation of AMPK.[28,29] Activity of respiratory chain complex I and AMP intracellular levels were monitored in DM1 MPCs after metformin treatment. Spectrometric measurements confirmed that progressive doses of metformin specifically inhibited the respiratory chain complex I in a dose-dependent manner up to a concentration of 10 mmol/l (**Figure 6a**). At 25 mmol/l, metformin also inhibited complexes II and IV and, to a lesser extent, complex V. This led to an increased AMP/ATP ratio, as determined by high-performance liquid chromatography (**Figure 6b**). The potential role of AMPK activation in alternative RNA splicing induced by metformin was tested by challenging DM1 MPCs with the AMPK activator AICAR (5-aminoimidazole-4-carboxamide 1-β-D-ribofuranoside, Acadesine, N^1-(β-D-ribofuranosyl)-5-aminoimidazole-4-carboxamide). Treatment of DM1 MPCs with 2 mmol/l AICAR for 24 hours promoted the downregulation of RBM3 (**Figure 6c**) together with changes in exon inclusions on five of the splicing events most regulated by metformin (*MDM4* exon 7, *GPCPD1* exon 5, *CCNL2* exon 7, *RAGE* exon 3, and *ZFAND1* exon 3) (**Figure 6d** and **Supplementary Table S3**). Strikingly, neither AICAR treatment nor downregulation of *RBM3* expression seemed to modulate the *INSR* exon 11 splicing in DM1 MPCs, suggesting that metformin might activate additional molecular pathways to mediate its global effect on alternative splicing (**Figure 6e**).

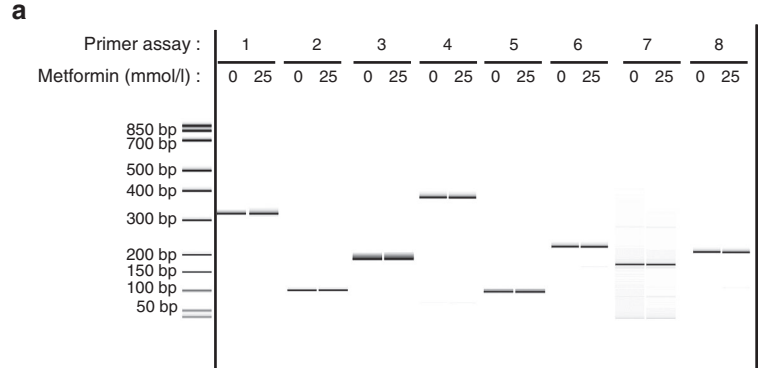
Impact of metformin on DM1-associated splicing defects in myoblasts

The potential interest of metformin in the pathological context of DM1 was evaluated by measuring the effect of metformin treatment on 20 splicing events that showed graded changes correlated with muscle strength in a cohort of 50 DM1 subjects.[30] This was accomplished by using primary cultures of myoblasts derived from two healthy individual and two distinct DM1 patients. In addition, splicing of *INSR* exon 11 and *TNNT2* exon 5, previously studied by the use of minigenes in DM1 MPCs (**Figure 3**), was analyzed as they are also expressed in cultured myoblasts. Metformin promoted changes in alternative splicing of exons ≥10% for six of these genes (**Figure 7** and **Supplementary Figure S4**). The effect of metformin was beneficial on *INSR* exon 11, *TNNT2* exon 5, *ATP2A1* exon 22, *DMD* exon 71, *DMD* exon 78, and *KIF13A* exon 32, as the isoform ratio shifted towards control values. The drug did not affect the 16 other splicing events that were tested (including those for which no splicing event is observed in absence of treatment in this cell type) (**Supplementary Figure S5**). The involvement of AMPK activation in the regulation of the six DM1 splicing defects modified by metformin was tested in one of the DM1 myoblast cultures. Experiments confirm a partial implication of the AMPK as AICAR treatment restored the splicing defects of *ATP2A1* exon 22 and *TNNT2* exon 5, while no modulations of *DMD* exon 78 and *INSR* exon 11 were observed (**Figure 7** and **Supplementary Figure S4**). Interestingly, AICAR promoted the inclusion of *DMD* exon 71. The splicing defect of *KIF13A* exon 32, not detected in these myoblasts, could not be analyzed. The efficacy of metformin treatment on DM1-associated splicing defects was next compared to pentamidine, a compound shown to revert some splicing defects associated with DM1.[6] Interestingly, a decreased expression of RBM3 is also observed after pentamidine treatment (**Supplementary Figure S6a,b**). Concordant with this observation, similar efficiency was observed between metformin and pentamidine with reference to their ability to restore the inclusion of *ATP2A1* exon 22, corresponding to the most affected splicing event in muscle biopsies of DM1 patients,[30] as well as on DMD exon 71 (**Figure 7**, **Supplementary Table S4**, and **Supplementary Figures S4** and **S6c**). However, in contrast to metformin treatment, no significant effect (>10%) of pentamidine was detected on *DMD* exon 78, *INSR* exon 11, and *TNNT2* exon 5.

In vivo effects of metformin on RNA alternative splicing

In order to explore the effects of metformin on alternative splicing at therapeutic concentrations in humans, we investigated patients currently being treated with metformin for Type 2 diabetes. To perform a clinical trial in which metformin would be temporarily replaced by another antidiabetic drug, namely sitagliptin, the lack of efficacy of sitagliptin was verified on *INSR* exon 11 alternative splicing in preliminary experiments carried

a



Transcript variant	Primer assay: 1	2	3	4	5	6	7	8	Biotype	Detected
RBM3-01	219								Processed transcript	No
RBM3-02	289								Processed transcript	No
RBM3-03 NM_006743	326					217		207	Protein coding	Yes
RBM3-04		91							Protein coding	Yes
RBM3-05	326						160		Protein coding	Yes
RBM3-06			209						Processed transcript	Yes
RBM3-07				377					Processed transcript	Yes
RBM3-08	742								Processed transcript	No
RBM3-09	595								Processed transcript	No
RBM3-10					101				Processed transcript	Yes
RBM3-11	488								Processed transcript	No
NM_1017430								100	NMD candidate	No
NM_1017431								476	NMD candidate	No

Size of the expexted amplicon (bp)

b

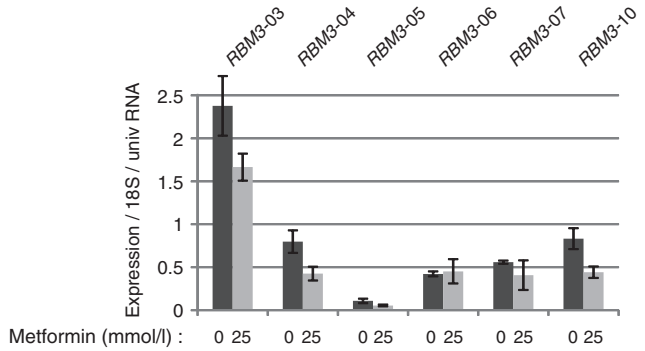


Figure 5 Analysis of the effect of metformin on *RBM3* alternative splicing. (**a**) Expression profile of 11 *RBM3* transcripts and 2 nonsense-mediated mRNA decay candidates in myotonic dystrophy type I (DM1) mesodermal precursor cells (MPCs) treated or not with 25 mmol/l of metformin for 48 hours. Primer assay corresponds to a specific pair of primers designed to match with distinct splice variants of *RBM3*. (**b**) Expression level of *RBM3* variants were quantified in DM1 MPC after a 48 hours treatment with 25 mmol/l metformin by reverse transcription–quantitative PCR. Primer assays that amplified one variant were used as followed: for *RBM3*-03, primer assay 6; for *RBM3*-04, primer assay 2; for *RBM3*-05, primer assay 7; for *RBM3*-06, primer assay 3; for *RBM3*-07, primer assay 4; for primer *RBM3*-10, primer assay 5. Data are represented as mean + SD. Expression level of each variant was normalized on 18S and on the expression in universal RNA. NMD, nonsense-mediated mRNA decay.

out *in vitro* on peripheral blood lymphocytes (PBLs) (**Supplementary Figure S7a,b**). Fifteen diabetic patients, who had been treated with a stable dose of metformin between 2.1 and 3 g/day for more than a year, were recruited in a study where metformin was replaced for 1 month by sitagliptin, and RNA alternative splicing was explored in PBLs (NCT 01349387). There was no change in blood glucose levels during the course of the study. Of the splicing events identified in response to metformin, we chose alternative splicing of *INSR* exon 11, which is expressed in most tissues. Because of its low level of

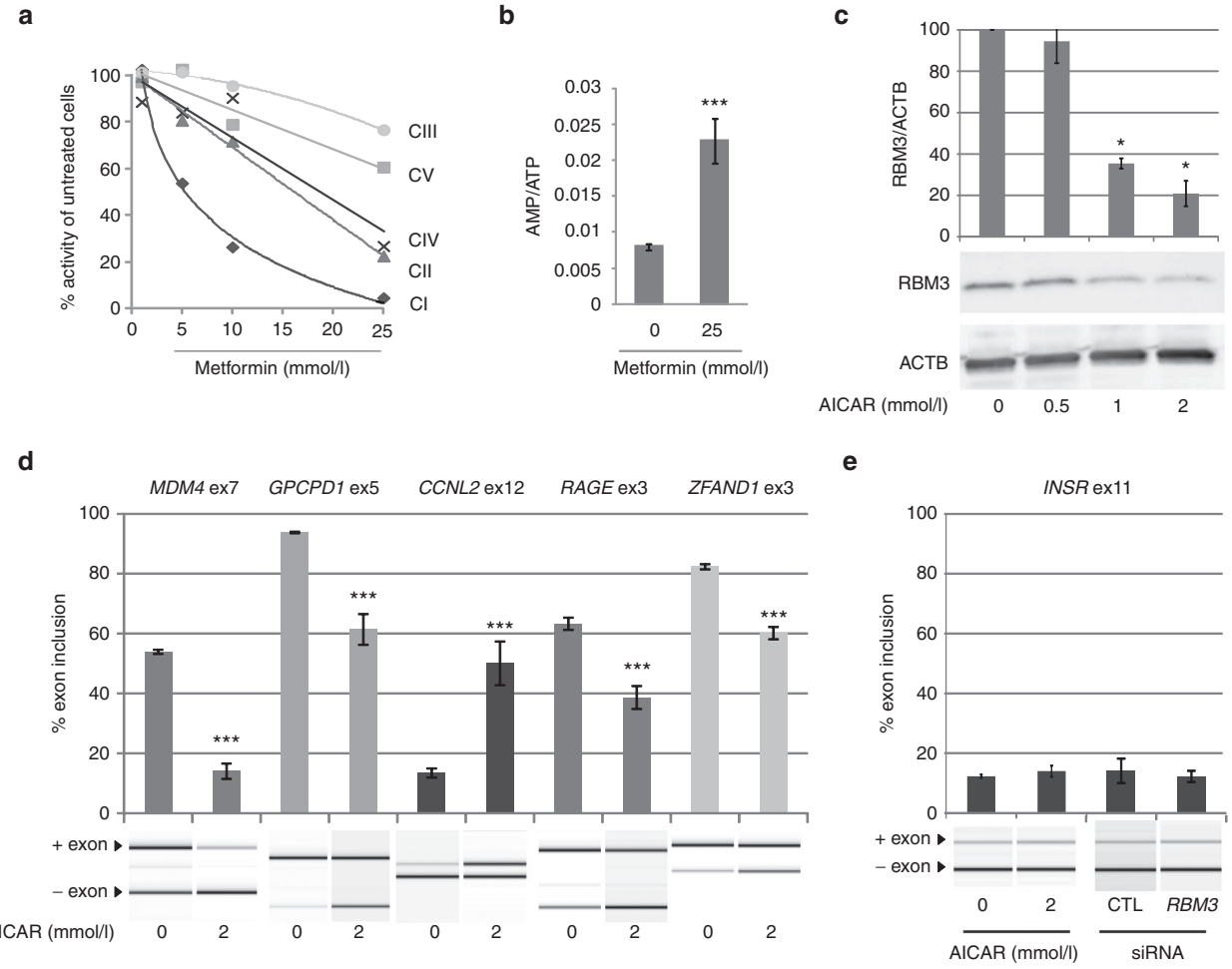


Figure 6 Intracellular modifications induced by metformin and involvement of AMPK in splicing control. (**a**) Activities of the different complexes of the respiratory chain were evaluated by spectrometry in myotonic dystrophy type I (DM1) mesodermal precursor cells (MPCs) treated for 48 hours with 5, 10, and 25 mmol/l metformin. Representative graph of the results obtained with three different samples. CI to CV represent the different complexes of the respiratory chain. (**b**) The effect of metformin on intracellular AMP/ATP content was quantified with high-performance liquid chromatography. Data (mean + SD) were analyzed with analysis of variance (ANOVA) and the Kruskal–Wallis *post hoc* test. (**c**) AICAR inhibits the expression of RBM3, analyzed by western blot. Data (mean + SD) were analyzed with ANOVA and the Kruskal–Wallis *post hoc*. (**d**) AICAR promotes changes in *MDM4* exon 7, *GPCPD1* exon 5, *CCNL2* exon 7, *RAGE* exon 3, and *ZFAND1* exon 3. (**e**) The alternative splicing of *INSR* exon 11, analyzed in DM1 MPCs by reverse transcription–PCR, is not modified by a 24-hour AICAR treatment or the downregulation of *RBM3* induced by the transfection of siRNA for 48 hours. Data (mean + SD) were analyzed with ANOVA and a Tukey–Kramer *post hoc* test. **P* < 0.05, ****P* < 0.001. AICAR, 5-aminoimidazole-4-carboxamide 1-β-D-ribofuranoside, Acadesine, N^1-(β-D-ribofuranosyl)-5-aminoimidazole-4-carboxamide).

expression in these cells, *INSR* +/– exon 11 transcripts were analyzed with quantitative PCR. Analysis confirmed that metformin triggered *INSR* exon 11 inclusion in this clinical setting (**Figure 8a**). Alternative splicing of *FAS (CD95)* exon 6 was analyzed in order to identify an additional splicing event that could be measured in PBLs from treated patients. *FAS* exon 6 exclusion was also affected by therapeutic doses of metformin in diabetic patients, giving rise to a shift from the anti- to the proapoptotic isoform of the protein (**Figure 8b**,**c**).

Discussion

The main result of this study is the demonstration that metformin, the antidiabetic biguanide, affects the alternative RNA splicing machinery. Our results point to a molecular mechanism that involves, at least partially, the activation of AMPK and modulation of the RBM3 RNA-binding protein. The demonstration that metformin modulates several splicing events *in vitro* and *in vivo*, including some altered in DM1, suggests that it would be worthwhile to evaluate the efficacy of metformin treatment in alleviating other symptoms than those related to insulin resistance.

The importance of gene regulation at the level of RNA transcripts by alternative splicing has been reported increasingly in fields such as development,[31] cancer,[32] metabolism,[33] and monogenic diseases.[1] Alternative RNA splicing thus opens new opportunities for therapeutic approaches.[34] However, it seems unlikely that selective modulation of a single specific splicing event could be carried out without generating concomitant adverse effects. However, this goal might be achieved with the use of marketed drugs whose biodistribution

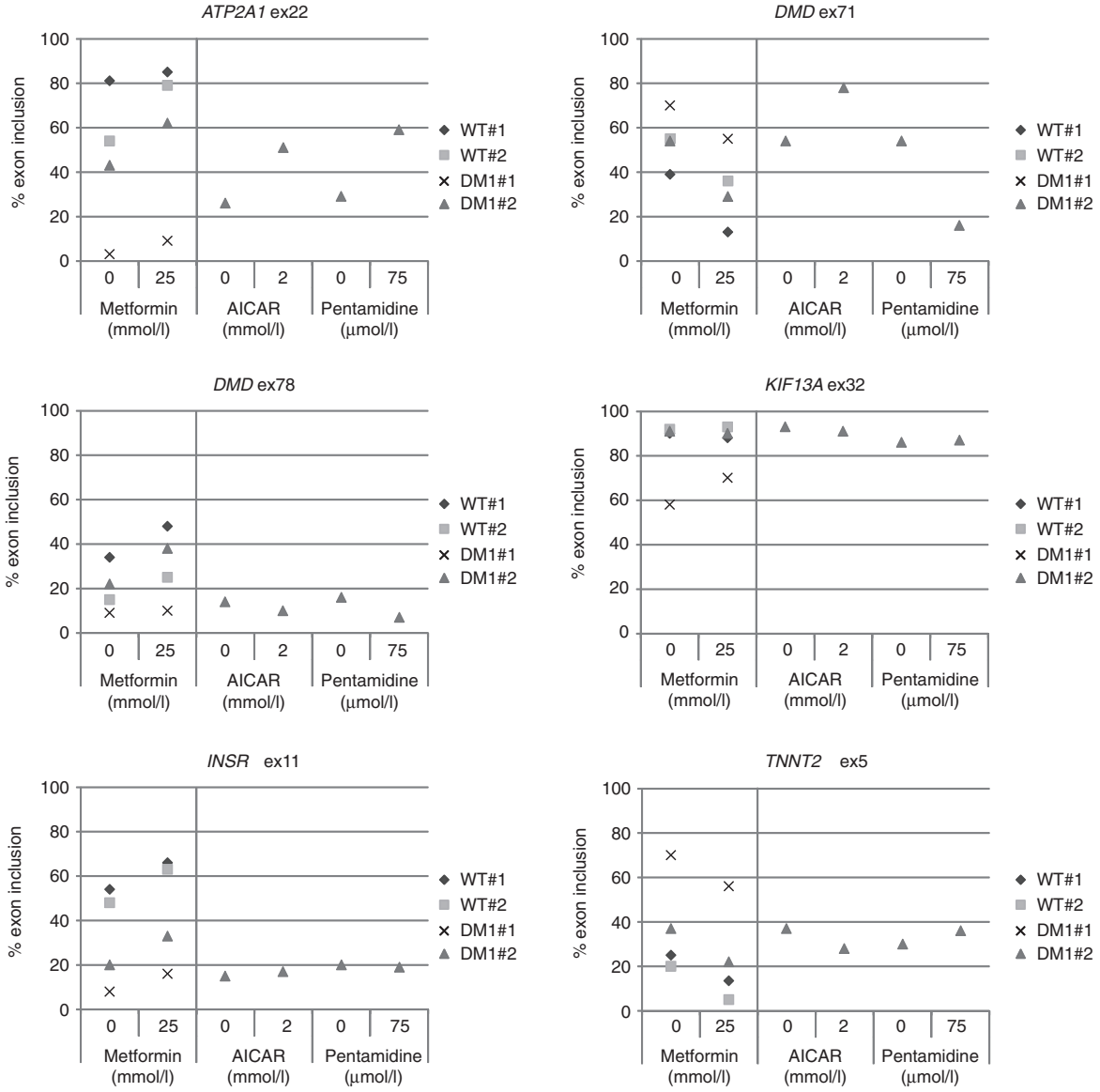


Figure 7 Metformin impacts several splicing defects associated with DM1 in mutated human myoblasts. Reverse transcription–PCR analysis of DM1-associated splicing defects in myoblasts from two DM1 patients and two healthy individuals treated for 48 hours with 25 mmol/l metformin or for 24 hours with 75 µmol/l pentamidine or AICAR 2 mmol/l. Representative data from two independent experiments are shown. AICAR, 5-aminoimidazole-4-carboxamide 1-β-D-ribofuranoside, Acadesine, N^1-(β-D-ribofuranosyl)-5-aminoimidazole-4-carboxamide; DM1, myotonic dystrophy type I; WT, wild type.

is known and which regulate alternative splicing.[4,5,35] These compounds could be reconsidered as modulators of alternative RNA splicing in addition to the effects on cellular targets for which they have been screened. It is worth mentioning that an analysis by exon array demonstrated that compounds such as clotrimazole, flunarizine, and chlorhexidine targeted different signal transduction pathways and caused distinct changes in alternative splicing of a number of genes.[4] This suggests that discrete targeting of the alternative splicing of specific genes associated with a particular disease may be possible with selected pharmacological agents.

In this context, we focused on metformin, which is indicated as a first-line oral therapy for treatment of hyperglycemia in individuals with Type 2 diabetes. Even though this drug has been in use for several decades, most of its cellular effects are still under investigation and new emerging effects, such as inhibition of cell proliferation, suggest its potential repurposing to treat cancer. Metformin was recently reported to selectively inhibit the translation of several RNA-binding proteins concomitantly with its blockade of cell proliferation.[7] Our results confirm that metformin treatment downregulates RBM3 in DM1 MPCs and induces splicing of a restricted set of primary transcripts. The cellular model we used allowed us to verify the involvement in this process of AMPK, the classical cellular target of metformin.[28,29] The characterization of metformin treatment in DM1 MPCs pointed to a signal associated with energy depletion and blockade of cell proliferation induced by inhibition of complex 1 of the mitochondria, ATP decrease, and activation

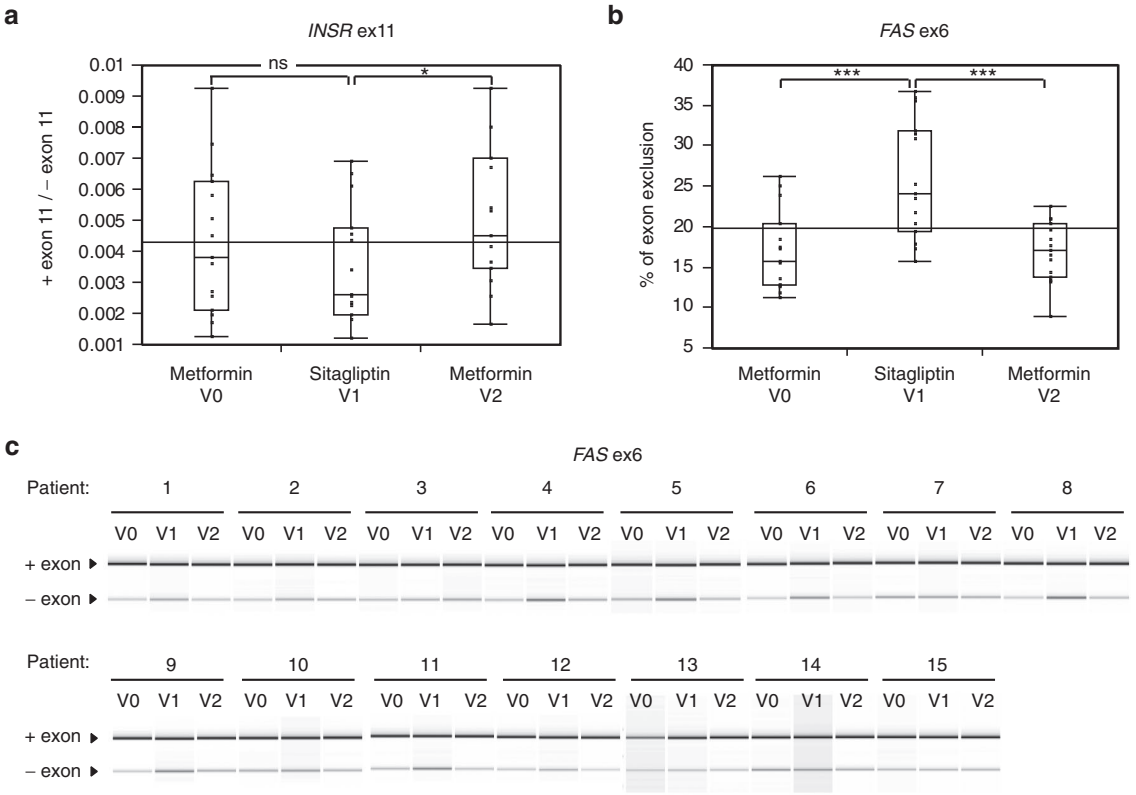


Figure 8 Regulation of alternative splicing of *INSR* exon 11 and *FAS* exon 6 by therapeutic doses of metformin can be followed *in vivo* in patients. (**a**) The *INSR* + exon 11 / – exon 11 ratio was monitored with reverse transcription–quantitative PCR (RT–qPCR) in peripheral blood lymphocytes from diabetic patients treated with metformin (V0), with sitagliptin instead of metformin for 1 month (V1) and 1 month after restarting metformin treatment (V2). (**b,c**) The same samples were also analyzed with RT–PCR to quantify the alternative splicing of *FAS* exon 6, which is also regulated by metformin. Data represent the individual responses in each group ($n = 15$ patients) and were analyzed with the Wilcoxon paired test. *$P < 0.05$, ***$P < 0.001$. NS, nonsignificant.

of the AMPK metabolic sensor. This reveals a molecular signature at the level of RNA transcripts that is associated with metabolic stress and similar to that identified for other genotoxic or oxidative stress inducers.[36,37] In parallel to these events, the absence of modulation of several DM1-associated splicing defects by the AMPK activator, AICAR, reveals the existence of additional molecular mechanisms by which metformin modulates the alternative splicing of certain primary transcripts. Metformin has been described as diminishing tyrosine kinase receptor signaling *in vitro* and *in vivo*.[38,39] These tyrosine kinase receptors include epidermal growth factor receptor, the signaling pathway of which controls *INSR* exon 11 inclusion through inhibition of *hnRNPA1* and *hnRNPA2B1* expression.[40] Whether such a mechanism is involved in other alternative splicing events regulated by metformin remains to be explored.

Among the regulated splicing events, we explored the impact of metformin treatment on those affected in DM1, because this drug is used to treat Type 2 diabetes in patients with DM1.[8] This well-tolerated drug could be an efficient way to alleviate DM1 missplicing in several organs affected by this multisystemic disease. It is commonly thought that most clinical manifestations of DM1 are linked with defects in alternative splicing due to the loss of MBNL1 function.[41] Accordingly, most attempts at finding treatments for this as yet incurable disease have focused on the release of MBNL1 from ribonucleoprotein intranuclear inclusions, and reversion of splicing defects has been observed with ribozymes,[42] antisense oligonucleotides,[43] and chemical compounds such as pentamidine.[6,44] Metformin is shown here to alter splicing through a different mechanism that targets the splicing machinery. The downregulation of RBM3 by metformin in DM1 but also wild-type MPCs reveals a mode of splicing regulation that is not specific to DM1. The impact of metformin on DM1-associated splicing defects could be in part defined by the overlap between the targets of RBM3 and those of MBNL1. Our results indicated that metformin is capable of alleviating several splicing defects in cells differentiated from pluripotent stem cells derived from a DM1-mutant embryo, as well as in myoblasts sampled from DM1 patients. To focus on DM1 splicing defects that would be therapeutically significant, we analyzed the impact of metformin treatment in DM1 myoblasts on 20 splicing defects identified in DM1 skeletal muscle tissue that were correlated with muscle weakness using a genome-wide approach.[30] Experiments confirmed the partial modulation of these splicings by metformin, including *ATP2A1* exon 22, identified as the most affected in correlation with muscle weakness in DM1 patients, and *INSR* exon 11 or *TTN* exon 5 that belong to the early transition splicing group that are strongly affected by DM1 (>30% shift of exon inclusion), yet not associated with muscle weakness ($r \leq 0.5$).[30] Within the *DMD* transcript, metformin enhanced the inclusion of exon 78 but also increased the DM1-associated skipping of exon 71.

A transcript lacking the *DMD* exon 71 is normally expressed in normal skeletal muscle but is overexpressed in DM1 patients. Immunoblot analysis shows no change in dystrophin protein expression in skeletal muscle between DM1 and non-DM individuals,[45] indicating that the functional impact of *DMD* exon 71 exclusion remains to be functionally tested. Notably, the skipping of *DMD* exon 71 is also observed in DM1 myoblasts in response to pentamidine treatment, indicating that treatments that restore MBNL1 expression also provide partial correction of the DM1-associated splicing defects. In parallel to the modulation of splicing, metformin could be of interest for DM1 as an activator of AMPK. AICAR treatment tested on muscle function in mdx mice[46] has been reported to promote significant improvements in disease phenotype (a gain in body and muscle weight, a decrease in muscle inflammation and in the number of fibers with central nuclei and an increase in fibers with peripheral nuclei), including an increase in overall behavioral activity and significant gains in forelimb and hind limb strength. Since metformin is commonly used to treat insulin resistance in DM1 patients at doses that were shown in the present study to induce shifts in transcript isoform ratios, we decided to investigate modulation of splicings by metformin as well as drug efficacy on several functional parameters in a clinical trial with DM1 patients (EudraCT number: 2013-001732-21). This study will determine the therapeutic potential of metformin to treat DM1 patients for aspects of their disorder other than insulin resistance.

Considering that metformin has been used for decades in millions of patients without major toxicity, one may consider targeting alternative splicing in order to obtain a therapeutic effect. Accordingly, a systematic search for the effects on alternative splicing of drugs that are already in current use may eventually allow clinicians to extend their indications to diseases in which a change in isoform ratios of specific genes may be therapeutically beneficial. For example, the two isoforms of *FAS* (CD95) have opposite effects, being either pro- or antiapoptotic,[47,48] and, in PBLs sampled from diabetic patients, treatment with metformin induced a shift from the antiapoptotic variant to the proapoptotic variant of CD95. This effect could influence FAS-mediated apoptosis, which could be relevant for Ewing and other sarcomas[49] or autoimmune lymphoproliferative syndromes resulting from the failure of *FAS* exon 6 inclusion.[50]

Materials and Methods

Reagents. Primers, probes, and siRNA sequences are listed in **Supplementary Table S5**. The *RBM3* siRNA came from Qiagen (Courtaboeuf, France). Sitagliptin was obtained from Januvia 100-mg tablets (MSD Merck Sharp & Dohme Ltd, Hoddesdon, UK). Metformin, AICAR (5-Aminoimidazole-4-carboxamide 1-β-D-ribofuranoside, Acadesine, N^1-(β-D-Ribofuranosyl)-5-aminoimidazole-4-carboxamide), pentamidine isethionate salt, cycloheximide, staurosporine, and ionomycin were obtained from Sigma. Primary antibodies used in this study were raised against SRSF1 (Clinisciences, Nanterre, France; LSB2340, 1/500), RBM3 (Abcam, Cambridge, UK; ab134946, 1/1,000), SFPQ (Abcam; ab117617, 1/500), RBM45 (Abcam; ab105770, 1/200), SRSF6 (Clinisciences; LS-B5712, 1/2,000), CELF1 (Millipore, Darmstadt, Germany; 05621, 1/2,000), Ki-67 (Millipore; MAB4190), and ACTB-peroxidase (Sigma-Aldrich, Saint-Louis, MO; A3854). Horseradish peroxidase-conjugated secondary antibodies used for western blot were goat anti-mouse IgG-horseradish peroxidase or goat anti-rabbit IgG-horseradish peroxidase (1:10,000; Amersham Bioscience, GE Healthcare, Saclay, France). MBNL1 was detected by the use of the MANDYS1 antibody, kindly provided by Prof. Glenn Morris (Center for Inherited Neuromuscular Disease, Oswestry, UK) and obtained from the MDA Monoclonal Antibody Resource.

Pluripotent stem cells culture. The two hESC lines used in this study came from the Department of Embryology and Genetics of the Vrije Universiteit, AZ-VUB Laboratory, Brussels, Belgium: the VUB03_DM1 (XX, passages 66–67) carrying the DM1 mutation (1,330 CTG repeats) and the VUB01_CTL (XY, passage 83) used as a control.[51] Human pluripotent stem cells were maintained on a layer of mitotically inactivated murine embryonic STO fibroblasts in Knockout Dulbecco's Modified Eagle's Medium supplemented with 20% knockout serum replacement, 1 mmol/l Glutamax, 1 mmol/l nonessential amino acids, 1% penicillin/streptomycin, 0.1% β-mercaptoethanol, and 5 ng/ml recombinant human FGF2 (all from Invitrogen, Carlsbad, CA). Medium was changed daily and cells were passaged every 5–7 days. Manual dissection was routinely used to passage the cells.

Differentiation of hESC lines in MPCs. MPCs were generated by differentiation from hESCs according to the protocol described previously by Marteyn *et al.*[19] MPCs derived from the VUB03_DM1 and VUB01_CTL hES cell lines were cultured on 0.1% gelatin-coated flasks and plates (Sigma-Aldrich) using Knockout Dulbecco's Modified Eagle's Medium (Invitrogen) supplemented with 20% fetal bovine serum (Eurobio, Les Ulis, France), 1 mmol/l Glutamax (Invitrogen), 1 mol/l nonessential amino acids (Invitrogen), and 0.1% β-mercaptoethanol (Invitrogen).

Culture of human myoblasts. Control and DM1 myoblasts were obtained from the Myobank in accordance with the French legislation on ethical rules (kindly provided by Dr. D. Furling). Two control myoblasts were originally isolated from the quadriceps of a 5-day-old infant (CHQ) and from a week 14 fetus (Me16). Two DM1 myoblasts were originally isolated from the quadriceps of a 11-day-old infant carrying more than 2,500 CTG (DM11) and from a week 14 fetus carrying 800 CTG (DM16). WT#1 and WT#2 correspond to Me16 and CHQ myoblasts while DM1#1 and DM1#2 match with DM11 and DM16 myoblasts. Cells were cultured on 0.1% gelatin-coated flasks and plates using Dulbecco's Modified Eagle's Medium-F12 + glutamax medium (Invitrogen) supplemented with 20% fetal bovine serum (Eurobio).

Culture of human PBLs. Freshly isolated wild-type and mutated PBLs, provided by Dr. Guillaume Bassez (CHU Henri Mondor, Creteil, France), were obtained from the Genethon DNA and Cells Bank (Evry, France). Cells were thawed and cultured in RPMI medium supplemented with 20% of fetal bovine serum (Eurobio) and 1% penicillin–streptomycin (Invitrogen)

according to cell bank instructions. Chemical treatments were performed once a day for 2 days. Freshly isolated PBLs from diabetic patients treated by metformin or sitagliptin were obtained from the CERITD (Evry, France) (NCT 01349387).

Measurement of sitagliptin activity and cell viability. Sitagliptin activity was measured *in vitro* using the luminescent DPPIV-Glo Protease Assay (Promega, Madison, WI) according the manufacturer's instructions. Viability of lymphocytes treated with a range of sitagliptin doses was monitored with the CellTiter-Glo assay (Promega) according the manufacturer's instructions.

Transfection of DNA constructs and siRNAs. MPCs were seeded in 24-well plates and transfected with 600 ng of plasmid, 0.6 µl of PLUS (Invitrogen) and 1.5 µl Lipofectamine LTX (Invitrogen). The RTB300 minigene used to analyze the splicing of exogenous human *cTNT* transcripts was kindly provided by Prof. TA Cooper (Baylor College of Medicine, Houston, TX). We constructed the minigene used to study *Clcn1* exon 7a splicing as described by Kino *et al.*[26] The genomic fragment covering exons 6 to 7 from mouse genomic DNA was PCR amplified using the *Clcn1* cloning primers described in **Supplementary Table S5**, cloned in the pCR-BluntII-TOPO vector (Invitrogen) using the BamHI/SalI restriction enzymes and then subcloned into the BglII-SalI site of pEGFP-C1 (Clontech, Mountain View, CA). For siRNA transfection, MPCs were seeded in 24-well plates and transfected with 10 nmol/l siRNA RBM3 (Qiagen) listed in **Supplementary Table S5** using 2.5 µl LipoRNAiMax (Invitrogen).

RNA sequencing library preparation and sequencing. Sequencing libraries were prepared according to the Illumina TruSeq Stranded mRNA Sample Prep Kit (according to manufacturer's protocol) (Illumina, San Diego, CA). A 2×101 bp paired-end sequencing was performed on the HiSeq2000 instrument, using half a lane per sample, to produce on average 80 million read pairs per sample (160 million sequences) with an average insert length of 130 bp. Trimmomatic,[52] Tophat2,[53] Picard suite (http://www.broadinstittute.github.io/picard), RNA-SeQC,[54] and in-house metrics were used to evaluate data quality.

RNA sequencing data analysis and identification of differential genes and splicing events. Reads were aligned using TopHat2 (v2.0.8[53]). TopHat2 was run with the assistance of gene annotations (Illumina's iGenomes based on EnsEMBL r70), which means that the alignment was performed in three steps: transcriptome mapping, genome mapping, and spliced mapping. The minimum and maximum intron lengths were also re-evaluated, respectively, to 30 and 1,200,000 to maximize the number of introns detected. The mate inner distances were set to their corresponding values. Alignment files in bam format were then filtered to removed poor mapping quality score (<10) and not primary alignments and read pairs with one single read mapped were filtered using samtools (v0.1.8[55]).

For the differential gene expression analysis, reads mapping to genes were first quantified using the HTSeq-count script provided by the HTSeq python package (v0.5.4[56]). The R/Bioconductor package DESeq2 (v1.4.5[57]) was then used to identify genes regulated by the drug treatment. A filter was then applied to DESeq2 results to select genes with an adjusted *P* value ≤0.05 and the mean of normalized counts ≥10.

For the alternative splicing analysis, reads crossing the exon–exon junction ("junction reads") were extracted from the read alignment files to detect the exon skipping events. In order to avoid spurious read alignments, we applied additional filters for the junction reads considered: no indels at the junction site, no hard clipping, and a minimal overlap of 4 bp over the junction site. FasterDB[58] gene and exon annotations were used as a guide to detect known and new exon skipping events. For each exon skipping event detected across all samples, junction reads corresponding to the inclusion of the exon and junction reads corresponding to the exclusion of the exon were quantified. The differential analysis was performed using KissDE, an R package developed as part of the KisSplice post-processing workflow.[59] KissDE works on pairs of variants for which read counts are available in each replicate of each condition and tests if a variant is enriched in one condition. Counts are modeled using a negative binomial distribution. KissDE fits a generalized linear model and tests for the effect of an interaction between the variant and the condition using a likelihood ratio test with a 5% false discovery rate to control for multiple testing. A percent splicing index (PSI or Ψ) value was then estimated for each sample as the ratio of inclusion junction reads to the sum of inclusion and exclusion junction reads. As the datasets are paired, the difference of Ψ values for each event (deltaPSI or $\Delta\Psi$) was calculated as the median of $\Delta\Psi$ values for each replicate. A filter was then applied on exon skipping events detected to select significant variants with an adjusted *P* value ≤0.05 and $\Delta\Psi$ value ≥10%.

Gene expression and splicing analysis by RT–PCR. Total RNA was extracted using the RNeasy Micro/Mini kit (Qiagen) and reverse transcribed using random hexamers and Superscript III Reverse Transcriptase kit (Invitrogen). For splicing analysis, PCR amplification was carried out with recombinant Taq DNA polymerase (Invitrogen) and the primers listed in **Supplementary Table S5**. The amplification was performed using a first step at 94 °C for 3 minutes followed by 30 cycles of 45 seconds at 94 °C, 30 seconds at 55 °C, 30 seconds at 72 °C, and finished with a final 10 minutes extension at 72 °C. The PCR products were quantified using the Bioanalyzer 2100 and DNA 1000 LabChip kit (Agilent, Santa Clara, CA). Primers used for splicing analysis in human myoblasts are described in Nakamori *et al.*[30] except those used to analyze *ATP2A1* exon 22 and *TNNT2* exon 5 splicings.

RBM3 gene expression and splicing analyses by quantitative PCR. Quantitative PCR reactions were carried out in 384-well plates using a QuantStudio 12K Flex Real-Time PCR System (Applied Biosystems, ThermoFisher Scientific, Illkirch Graffenstaden, France) with Power SYBR Green 2× Master Mix (Life Technologies, ThermoFisher Scientific, Illkirch Graffenstaden, France), 0.5 µl of cDNA, and 100 nmol/l of primers (Invitrogen) in a final volume of 10 µl. Detailed information on the primers sequences is provided in **Supplementary Table S5**. The relative expression level of each gene was calculated with the method described by Pfaffl.[60] A precise description of samples preparation and

experiment procedure are compiled in **Supplementary Table S6**. Data were expressed as mean ± SD.

Protein extraction and western blot analysis. Cells were homogenized in radioimmunoprecipitation assay buffer (Sigma-Aldrich) containing 1% protease inhibitors (Sigma-Aldrich) and 10% phosphatase inhibitors (Roche, Paris, France). After electrophoresis on 4–12% Nu-PAGE Bis-Tris gels (Invitrogen) under reducing conditions, proteins were transferred to nitrocellulose membranes (Invitrogen), blocked with phosphate-buffered saline (PBS) containing 0.1% Tween-20 and 5% bovine serum albumin (BSA) or 5% nonfat dry milk, depending on the primary antibody used, and incubated overnight with the primary antibody diluted in PBS containing 0.1% Tween-20 and 5% BSA or 5% nonfat dry milk. Membranes were then incubated for 1 hour with the corresponding secondary antibody and immunoreactive protein bands were detected by ECL Plus detection reagents (Amersham Bioscience) according to the manufacturer's protocol using an ImageQuant CDD camera (GE Healthcare).

Enzymatic activities. Respiratory chain enzyme activities were spectrophotometrically measured using a Cary 50 UV–visible spectrophotometer (Varian, Les Ulis, France) as described by Bénit *et al.*[61] Mitochondrial substrate oxidation was polarographically estimated using a Clark oxygen electrode (Hansatech Instruments, Norfolk, England) in a magnetically stirred 250-µl chamber maintained at 37 °C in 250 µl of a respiratory medium consisting of 0.3 mol/l mannitol, 5 mmol/l KCl, 5 mmol/l $MgCl_2$, 10 mmol/l phosphate buffer (pH 7.2), and 1 mg/ml BSA, plus substrates or inhibitors as described by Rustin *et al.*[62] Protein concentration was measured according to the Bradford assay.

Viability, cytotoxicity, and caspase assay. MPCs DM1 were seeded at 5,000 cells/well in 96-well plate and were treated with dose range of metformin for 48 hours or dose range of ionomycin and staurosporine for 24 hours. Viability, cytotoxicity, and apoptosis events were assessed using the ApoTox-Glo Triplex Assay (Promega). After incubation of cells with the "Viability/Cytotoxicity reagent" for 50 minutes at 37 °C, the resulting cell viability and cytotoxicity fluorescences were measured respectively at 400Ex/505Em and 485Ex/520Em using the CLARIOstar microplate reader (BMG LABTECH, Champigny-sur-Marne, France). Cells were then incubated with the "Caspase-Glo 3/7 reagent" for 30 minutes at room temperature in dark, and caspase activation (a hallmark of apoptosis) was determined with luminescence measurement using the CLARIOstar microplate reader (BMG LABTECH).

Ki-67 proliferation assay. DM1 MPCs were treated with metformin dose range for 48 hours, daily repeated. After treatment, cells were fixed with 4% paraformaldehyde in PBS for 15 minutes at room temperature and incubated overnight at 4 °C with Ki-67 antibody diluted in PBS solution with 0.1% BSA and 0.3% Triton. After three washings in PBS, cells were incubated for 1 hour at room temperature with Alexa Fluor 647 goat anti-mouse IgG (ref. A21235; 1:1,000; Invitrogen) and Hoechst (ref. H3570; 1:3,000; Invitrogen) diluted in the same blocking solution as previously. After three washings in PBS, percentage of cells in proliferation was assessed by counting Ki-67–positive nuclei number using a Cellomics Arrayscan automated microscope (Thermo Scientific, Hudson, NH).

Metforgene clinical trial for the INSR exon 11 splicing monitoring in diabetic patients. An interventional clinical trial was promoted by CERITD in the Centre Hospitalier Sud Francilien (Corbeil-Essonnes, France) (NCT 01349387) to investigate whether a treatment with metformin in patients with Type 2 diabetes had an effect on *INSR* exon 11 alternative splicing of the insulin receptor. During their visit of consultation on the follow-up to the Type 2 diabetes, 15 patients were selected on the basis of active metformin treatment at a dose greater than or equal to 1,400 mg/day. After inclusion in the study to day 0, metformin treatment will be interrupted between day 1 and day 30, replaced by Januvia 100 mg/day dose, and then resumed at day 31. Patients had to achieve a 10 ml blood sample at day 0, day 30, and 1 month after metformin retreatment. Blood samples were processed by Ficoll gradient centrifugation to isolate the circulating leukocytes. Total RNA was extracted using the RNeasy Micro/Mini kit (Qiagen) and reverse transcribed using random hexamers and Superscript III Reverse Transcriptase kit (Invitrogen). Expressions of *INSR* +/– exon 11 transcripts and 18S were monitored with TaqMan gene expression assays using the primers and MGB probes described in **Supplementary Table S5** and TaqMan Gene Expression Master Mix (Applied Biosystems) using the 7900HT Fast Real-Time PCR System (Applied Biosystems). The method described by Pfaffl[60] was used to determine the relative expression level of each gene. *FAS* exon 6 alternative splicing was additionally tested by RT–PCR. PCR amplification was carried out with recombinant Taq DNA polymerase (Invitrogen) and the primers listed in **Supplementary Table S5**. The amplification was performed using a first step at 94 °C for 3 minutes followed by 30 cycles of 45 seconds at 94 °C, 30 seconds at 55 °C, 30 seconds at 72 °C, and finished with a final 10 minutes extension at 72 °C. The PCR products were quantified using the Bioanalyzer 2100 and DNA 1000 LabChip kit (Agilent). Statistics were computed using JMP9 software (SAS, Cary, NC). Statistical differences were determined with a Wilcoxon paired test. Differences between groups were considered significant when $P < 0.05$ ($^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

Statistical analysis. Statistics were computed in JMP using *P* values. Values are reported as mean and SD. Differences between groups were considered significant when $P < 0.05$ ($^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$). According the size of the experiment, samples parametric (ANOVA and *post hoc* tests) or nonparametric tests were chosen.

Supplementary Material

Figure S1. Western blot analysis of RBM3, SRSF1, SRSF6, RBM45 and SFPQ expressions in wild type or DM1 MPCs in response to metformin treatment.
Figure S2. Heatmap representation of the splicing events modulated by metformin in DM1 MPCs and analysis of RBM3 RNA-binding protein involvement in this regulation.
Figure S3. Analysis of RBM3 transcript variants that are candidate to the non sense mediated decay in response to metformin and cycloheximide treatments in DM1 MPCs.

Figure S4. RT-PCR detection of 6 DM1 associated splicing defects in myoblasts from 2 non affected individuals or 2 DM1 patients, treated with metformin, pentamidine or AICAR.
Figure S5. Metformin does not impact the alternative splicing of 16 splicing defects in DM1 mutated myoblasts (DM16) treated for 48 hours with a range of dose of metformin.
Figure S6. Impact of pentamidine on RBM3 expression and DM1 associated splicing defects in DM1 human myoblasts.
Figure S7. In vitro evaluation of metformin and sitagliptin treatments on *INSR* exon 11 splicing in peripheral blood lymphocytes.
Table S1. List of genes modulated by 10mM metformin treatment for 48 hours in DM1 MPCS.
Table S2. List of genes modulated by 25mM metformin treatment for 48 hours in DM1 MPCS.
Table S3. List of splicing events modulated by 10 mM and 25 mM metformin treatments for 48 hours in DM1 MPCS.
Table S4. Effect of 25 mM metformin, 75 µM pentamidine and 2 mM AICAR on the regulation of alternative splicings altered in the DM1 mutated human myoblasts DM16.
Table S5. Human primers, probes and siRNA sequences.
Table S6. Detailed procedures of reverse transcription-quantitative PCR experiments according to the MIQE Guideli.

Acknowledgments. We thank Laetitia Aubry (INSERM/UEVE UMR 861, Evry, France) for discussions. The authors also thank Karen Sermon (Department of Embryology and Genetics, Vrije Universiteit Brussel, Brussels, Belgium) for providing embryonic stem cell lines, Thomas Andy Cooper (Baylor College of Medicine, Houston, TX) for *cTNT* minigenes, Odile Jouy and Marie-Hélène Petit for organizing the Metforgene clinical trial on diabetic patients (NCT 01349387) and collection of blood samples, Morgane Gauthier for the purification of primary human myoblasts, Safa Saker (Genethon, Evry, France), and the Genethon DNA and Cell Bank for processing patients' blood samples. I-Stem is part of the Biotherapies Institute for Rare Diseases (BIRD) supported by the Association Française contre les Myopathies (AFM-Téléthon). This work was supported in part by INSERM, AFM-Téléthon (Association Française des Myopathes), and additional grants from the European Commission (STEM-HD, FP6), the Labex REVIVE, and DIM Stem Pôle. P.B. and P.R. were supported by AMMi and ANR. The authors declare that there are no competing financial interests in relation to the work described.

1. Cooper, TA, Wan, L and Dreyfuss, G (2009). RNA and disease. *Cell* **136**: 777–793.
2. Keren, H, Lev-Maor, G and Ast, G (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**: 345–355.
3. Oana, K, Oma, Y, Suo, S, Takahashi, MP, Nishino, I, Takeda, S *et al.* (2013). Manumycin A corrects aberrant splicing of Clcn1 in myotonic dystrophy type 1 (DM1) mice. *Sci Rep* **3**: 2142.
4. Younis, I, Berg, M, Kaida, D, Dittmar, K, Wang, C and Dreyfuss, G (2010). Rapid-response splicing reporter screens identify differential regulators of constitutive and alternative splicing. *Mol Cell Biol* **30**: 1718–1728.
5. Stoilov, P, Lin, CH, Damoiseaux, R, Nikolic, J and Black, DL (2008). A high-throughput screening strategy identifies cardiotonic steroids as alternative splicing modulators. *Proc Natl Acad Sci USA* **105**: 11218–11223.
6. Warf, MB, Nakamori, M, Matthys, CM, Thornton, CA and Berglund, JA (2009). Pentamidine reverses the splicing defects associated with myotonic dystrophy. *Proc Natl Acad Sci USA* **106**: 18551–18556.
7. Larsson, O, Morita, M, Topisirovic, I, Alain, T, Blouin, MJ, Pollak, M *et al.* (2012). Distinct perturbation of the translatome by the antidiabetic drug metformin. *Proc Natl Acad Sci USA* **109**: 8977–8982.
8. Kouki, T, Takasu, N, Nakachi, A, Tamanaha, T, Komiya, I and Tawata, M (2005). Low-dose metformin improves hyperglycaemia related to myotonic dystrophy. *Diabet Med* **22**: 346–347.
9. Ho, TH, Charlet-B, N, Poulos, MG, Singh, G, Swanson, MS and Cooper, TA (2004). Muscleblind proteins regulate alternative splicing. *EMBO J* **23**: 3103–3112.
10. Kanadia, RN, Johnstone, KA, Mankodi, A, Lungu, C, Thornton, CA, Esson, D *et al.* (2003). A muscleblind knockout model for myotonic dystrophy. *Science* **302**: 1978–1980.
11. Miller, JW, Urbinati, CR, Teng-Umnuay, P, Stenberg, MG, Byrne, BJ, Thornton, CA *et al.* (2000). Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J* **19**: 4439–4448.
12. Brook, JD, McCurrach, ME, Harley, HG, Buckler, AJ, Church, D, Aburatani, H *et al.* (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **69**: 385.
13. Fu, YH, Pizzuti, A, Fenwick, RG Jr, King, J, Rajnarayan, S, Dunne, PW *et al.* (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**: 1256–1258.
14. Mahadevan, M, Tsilfidis, C, Sabourin, L, Shutler, G, Amemiya, C, Jansen, G *et al.* (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255**: 1253–1255.
15. Kuyumcu-Martinez, NM, Wang, GS and Cooper, TA (2007). Increased steady-state levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. *Mol Cell* **28**: 68–78.
16. Fugier, C, Klein, AF, Hammer, C, Vassilopoulos, S, Ivarsson, Y, Toussaint, A *et al.* (2011). Misregulated alternative splicing of BIN1 is associated with T tubule alterations and muscle weakness in myotonic dystrophy. *Nat Med* **17**: 720–725.
17. Charlet-B, N, Savkur, RS, Singh, G, Philips, AV, Grice, EA and Cooper, TA (2002). Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing. *Mol Cell* **10**: 45–53.
18. Mankodi, A, Takahashi, MP, Jiang, H, Beck, CL, Bowers, WJ, Moxley, RT *et al.* (2002). Expanded CUG repeats trigger aberrant splicing of ClC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy. *Mol Cell* **10**: 35–44.
19. Marteyn, A, Maury, Y, Gauthier, MM, Lecuyer, C, Vernet, R, Denis, JA *et al.* (2011). Mutant human embryonic stem cells reveal neurite and synapse formation defects in type 1 myotonic dystrophy. *Cell Stem Cell* **8**: 434–444.
20. Denis, JA, Gauthier, M, Rachdi, L, Aubert, S, Giraud-Triboult, K, Poydenot, P *et al.* (2013). mTOR-dependent proliferation defect in human ES-derived neural stem cells affected by myotonic dystrophy type 1. *J Cell Sci* **126**(Pt 8): 1763–1772.
21. Gauthier, M, Marteyn, A, Denis, JA, Cailleret, M, Giraud-Triboult, K, Aubert, S *et al.* (2013). A defective Krab-domain zinc-finger transcription factor contributes to altered myogenesis in myotonic dystrophy type 1. *Hum Mol Genet* **22**: 5188–5198.
22. Holt, I, Mittal, S, Furling, D, Butler-Browne, GS, Brook, JD and Morris, GE (2007). Defective mRNA in myotonic dystrophy accumulates at the periphery of nuclear splicing speckles. *Genes Cells* **12**: 1035–1048.
23. Whelan, RS, Konstantinidis, K, Wei, AC, Chen, Y, Reyna, DE, Jha, S *et al.* (2012). Bax regulates primary necrosis through mitochondrial dynamics. *Proc Natl Acad Sci USA* **109**: 6566–6571.
24. Chinnaiyan, AM, Tepper, CG, Seldin, MF, O'Rourke, K, Kischkel, FC, Hellbardt, S *et al.* (1996). FADD/MORT1 is a common mediator of CD95 (Fas/APO-1) and tumor necrosis factor receptor-induced apoptosis. *J Biol Chem* **271**: 4961–4965.
25. Philips, AV, Timchenko, LT and Cooper, TA (1998). Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science* **280**: 737–741.
26. Kino, Y, Washizu, C, Oma, Y, Onishi, H, Nezu, Y, Sasagawa, N *et al.* (2009). MBNL and CELF proteins regulate alternative splicing of the skeletal muscle chloride channel CLCN1. *Nucleic Acids Res* **37**: 6477–6490.
27. Durand, S, Cougot, N, Mahuteau-Betzer, F, Nguyen, CH, Grierson, DS, Bertrand, E *et al.* (2007). Inhibition of nonsense-mediated mRNA decay (NMD) by a new chemical molecule reveals the dynamic of NMD factors in P-bodies. *J Cell Biol* **178**: 1145–1160.
28. El-Mir, MY, Nogueira, V, Fontaine, E, Avéret, N, Rigoulet, M and Leverve, X (2000). Dimethylbiguanide inhibits cell respiration via an indirect effect targeted on the respiratory chain complex I. *J Biol Chem* **275**: 223–228.
29. Zhou, G, Myers, R, Li, Y, Chen, Y, Shen, X, Fenyk-Melody, J *et al.* (2001). Role of AMP-activated protein kinase in mechanism of metformin action. *J Clin Invest* **108**: 1167–1174.
30. Nakamori, M, Sobczak, K, Puwanant, A, Welle, S, Eichinger, K, Pandya, S *et al.* (2013). Splicing biomarkers of disease severity in myotonic dystrophy. *Ann Neurol* **74**: 862–872.
31. Gabut, M, Samavarchi-Tehrani, P, Wang, X, Slobodeniuc, V, O'Hanlon, D, Sung, HK *et al.* (2011). An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* **147**: 132–146.
32. Anczuków, O, Rosenberg, AZ, Akerman, M, Das, S, Zhan, L, Karni, R *et al.* (2012). The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nat Struct Mol Biol* **19**: 220–228.
33. Pihlajamäki, J, Lerin, C, Itkonen, P, Boes, T, Floss, T, Schroeder, J *et al.* (2011). Expression of the splicing factor gene SFRS10 is reduced in human obesity and contributes to enhanced lipogenesis. *Cell Metab* **14**: 208–218.
34. Naryshkin, NA, Weetall, M, Dakka, A, Narasimhan, J, Zhao, X, Feng, Z *et al.* (2014). Motor neuron disease. SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy. *Science* **345**: 688–693.
35. Anderson, ES, Lin, CH, Xiao, X, Stoilov, P, Burge, CB and Black, DL (2012). The cardiotonic steroid digitoxin regulates alternative splicing through depletion of the splicing factors SRSF3 and TRA2B. *RNA* **18**: 1041–1049.
36. Dutertre, M, Sanchez, G, De Cian, MC, Barbier, J, Dardenne, E, Gratadou, L *et al.* (2010). Cotranscriptional exon skipping in the genotoxic stress response. *Nat Struct Mol Biol* **17**: 1358–1366.

37. Vivarelli, S, Lenzken, SC, Ruepp, MD, Ranzini, F, Maffioletti, A, Alvarez, R *et al.* (2013). Paraquat modulates alternative pre-mRNA splicing by modifying the intracellular distribution of SRPK2. *PLoS One* **8**: e61980.
38. Quinn, BJ, Dallos, M, Kitagawa, H, Kunnumakkara, AB, Memmott, RM, Hollander, MC *et al.* (2013). Inhibition of lung tumorigenesis by metformin is associated with decreased plasma IGF-I and diminished receptor tyrosine kinase signaling. *Cancer Prev Res (Phila)* **6**: 801–810.
39. Iglesias, DA, Yates, MS, van der Hoeven, D, Rodkey, TL, Zhang, Q, Co, NN *et al.* (2013). Another surprise from Metformin: novel mechanism of action via K-Ras influences endometrial cancer response to therapy. *Mol Cancer Ther* **12**: 2847–2856.
40. Chettouh, H, Fartoux, L, Aoudjehane, L, Wendum, D, Clapéron, A, Chrétien, Y *et al.* (2013). Mitogenic insulin receptor-A is overexpressed in human hepatocellular carcinoma due to EGFR-mediated dysregulation of RNA splicing factors. *Cancer Res* **73**: 3974–3986.
41. Du, H, Cline, MS, Osborne, RJ, Tuttle, DL, Clark, TA, Donohue, JP *et al.* (2010). Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. *Nat Struct Mol Biol* **17**: 187–193.
42. Wheeler, TM, Leger, AJ, Pandey, SK, MacLeod, AR, Nakamori, M, Cheng, SH *et al.* (2012). Targeting nuclear RNA for *in vivo* correction of myotonic dystrophy. *Nature* **488**: 111–115.
43. Leger, AJ, Mosquea, LM, Clayton, NP, Wu, IH, Weeden, T, Nelson, CA *et al.* (2013). Systemic delivery of a Peptide-linked morpholino oligonucleotide neutralizes mutant RNA toxicity in a mouse model of myotonic dystrophy. *Nucleic Acid Ther* **23**: 109–117.
44. Coonrod, LA, Nakamori, M, Wang, W, Carrell, S, Hilton, CL, Bodner, MJ *et al.* (2013). Reducing levels of toxic RNA with small molecules. *ACS Chem Biol* **8**: 2528–2537.
45. Nakamori, M, Kimura, T, Fujimura, H, Takahashi, MP and Sakoda, S (2007). Altered mRNA splicing of dystrophin in type 1 myotonic dystrophy. *Muscle Nerve* **36**: 251–257.
46. Jahnke, VE, Van Der Meulen, JH, Johnston, HK, Ghimbovschi, S, Partridge, T, Hoffman, EP *et al.* (2012). Metabolic remodeling agents show beneficial effects in the dystrophin-deficient mdx mouse model. *Skelet Muscle* **2**: 16.
47. Cheng, J, Zhou, T, Liu, C, Shapiro, JP, Brauer, MJ, Kiefer, MC *et al.* (1994). Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science* **263**: 1759–1762.
48. Cascino, I, Fiucci, G, Papoff, G and Ruberti, G (1995). Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. *J Immunol* **154**: 2706–2713.
49. Paronetto, MP, Bernardis, I, Volpe, E, Bechara, E, Sebestyén, E, Eyras, E *et al.* (2014). Regulation of FAS exon definition and apoptosis by the Ewing sarcoma protein. *Cell Rep* **7**: 1211–1226.
50. Liu, C, Cheng, J and Mountz, JD (1995). Differential expression of human Fas mRNA species upon peripheral blood mononuclear cell activation. *Biochem J* **310 (Pt 3)**: 957–963.
51. Mateizel, I, De Temmerman, N, Ullmann, U, Cauffman, G, Sermon, K, Van de Velde, H *et al.* (2006). Derivation of human embryonic stem cell lines from embryos obtained after IVF and after PGD for monogenic disorders. *Hum Reprod* **21**: 503–511.
52. Bolger, AM, Lohse, M and Usadel, B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.
53. Kim, D, Pertea, G, Trapnell, C, Pimentel, H, Kelley, R and Salzberg, SL (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.
54. DeLuca, DS, Levin, JZ, Sivachenko, A, Fennell, T, Nazaire, MD, Williams, C *et al.* (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**: 1530–1532.
55. Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N *et al.*; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
56. Anders, S, Pyl, PT and Huber, W (2014). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.
57. Anders, S and Huber, W (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.
58. Mallinjoud, P, Villemin, JP, Mortada, H, Polay Espinoza, M, Desmet, FO, Samaan, S *et al.* (2014). Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome Res* **24**: 511–521.
59. Sacomoto, GA, Kielbassa, J, Chikhi, R, Uricaru, R, Antoniou, P, Sagot, MF *et al.* (2012). KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* **13** (suppl. 6): S5.
60. Pfaffl, M (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**: e45.
61. Bénit, P, Goncalves, S, Philippe Dassa, E, Brière, JJ, Martin, G and Rustin, P (2006). Three spectrophotometric assays for the measurement of the five respiratory chain complexes in minuscule biological samples. *Clin Chim Acta* **374**: 81–86.
62. Rustin, P, Chretien, D, Bourgeron, T, Gérard, B, Rötig, A, Saudubray, JM *et al.* (1994). Biochemical and molecular investigations in respiratory chain deficiencies. *Clin Chim Acta* **228**: 35–51.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

Supplementary Information accompanies this paper on the Molecular Therapy–Nucleic Acids website (http://www.nature.com/mtna)

II Exon ontologie

L'article ci-dessous a été soumis à *Genome Research* en juillet 2016.

Le principe de l'exon ontologie a été présenté dans la section III.A des résultats. Cet article présente l'outil ainsi qu'un cas d'utilisation de cet outil et montre l'intérêt d'une telle méthode pour accélérer la caractérisation des conséquences biologiques provenant de variations d'épissage.

Ma principale participation dans ce projet fut l'analyse de données RNAseq publics avec FaRLine. La première analyse fut de comparer des lignées cellulaires épithéliales à des lignées cellulaires fibroblastiques. Les données RNA-Seq de 7 lignées fibroblastiques (mésenchymales) normales et 3 lignées épithéliales normales ont été récupérées sur ENCODE. La seconde analyse a consisté à comparer deux sous-types de lignées de cancer du sein : des lignées de type Luminal (dites "epithelial-like") et des lignées de type Claudin-Low (dites "mesenchymal-like"). Ces données proviennent d'une étude à large échelle sur le cancer du sein [Daemen et al., 2013]. De ces comparaisons, une liste d'exons régulés entre les cellules épithéliales et fibroblastiques a été établie. Ces événements d'épissage ont été validés par RT-PCR avec un très bon taux de validation et une très bonne corrélation entre les $\Delta\Psi$ des RNA-Seq et des RT-PCR.

L'outil d'exon ontologie a ensuite été utilisé pour analyser cette liste d'exons régulés. Cela a permis de mettre en évidence qu'un grand nombre d'exons qui code pour des domaines protéiques contenant des sites de phosphorylation (validés expérimentalement) sont régulés entre les cellules de type épithéliale et celles de type mésenchymateuse. La recherche de potentielles séquences consensus de ces sites de phosphorylation a permis de montrer pour la première fois le rôle des événements d'épissage régulés par ESRP dans la voie de signalisation d'AKT dans les cellules épithéliales. L'approche d'exon ontologie a permis de découvrir des propriétés des protéines qui apparaissaient ensemble dans des exons alternatifs. Cela a permis de mettre en évidence un lien entre épissage alternatif et le processus d'autophagie.

J'ai également participé au développement de la méthode statistique pour mesurer l'enrichissement des termes de l'exon ontologie. Un schéma explicatif de cette méthode peut être trouvée dans la figure 2 A de l'article.

Exon Ontology: Functional Genomics

At Exon Level Resolution

Léon-Charles Tranchevent[1], Fabien Aubé[1], Louis Dulaurier[1], Clara Benoit-Pilven[1], Amandine Rey[1], Arnaud Poret[1], Emilie Chautard[2], Hussein Mortada[1], François-Olivier Desmet[1], Fatima Zahra Chakrama[1], Maira Alejandra Moreno-Garcia[1], Evelyne Goillot[3], Stéphane Janczarski[1], Franck Mortreux[1], Cyril F. Bourgeois[1,$], Didier Auboeuf[1,$,*]

[1.] Univ Lyon, ENS de Lyon, Univ 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.] Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, UMR CNRS 5558, INRIA Erable, Lyon, France

[3.] Institut NeuroMyoGène, CNRS UMR 5310, INSERM U1217, Université de Lyon, Université Claude Bernard Lyon 1, Lyon, France

[$] equal contribution

*Corresponding author: Didier Auboeuf, Laboratory of Biology and Modelling of the Cell, ENS de Lyon, 69007 Lyon, France. Didier.auboeuf@inserm.fr

Abstract

Transcriptomic genome-wide analyses demonstrate massive variations of alternative splicing in many physiological and pathological situations. One major challenge is now to establish the biological contribution of alternative splicing variations in physiological- or pathological-associated cellular phenotypes. Toward this end, we developed a computational suite, named “Exon Ontology” based on Exon Ontology terms corresponding to well characterized protein features organized in an ontology tree. “Exon Ontology” is conceptually similar to “Gene Ontology”-based approaches but focuses on exon level functional features instead of gene level functional annotations. “Exon Ontology” describes the protein features encoded by a selected list of exons and looks for potential Exon Ontology term enrichment. Applying “Exon Ontology” to a set of exons that are differentially spliced between epithelial and mesenchymal cells, we experimentally demonstrate that “Exon Ontology” discovers specific protein features and functions regulated by alternative splicing. We also show that “Exon Ontology” unravels biological processes that depend on suites of co-regulated alternative exons, as we uncovered a role of epithelial-enriched splicing factors in the AKT signaling pathway and of mesenchymal-enriched splicing factors in driving splicing events impacting on biological processes, like autophagy. Freely available on the web (http://fasterdb.ens-lyon.fr/ExonOntology/), Exon Ontology is the first resource providing computational support for the research community to predict the systems-biological consequences of alternative splicing.

Introduction

Alternative splicing is a major step in the gene expression process leading to the production from one single gene of different transcripts with different exon content (or alternative splicing variants). This mechanism is the rule as 95% of human genes produce at least two splicing variants (Nilsen and Graveley 2010; de Klerk and t Hoen 2015; Lee and Rio 2015). Alternative splicing decisions rely on splicing factors binding on pre-mRNA molecules more or less close to splicing sites and regulating their recognition by the spliceosome (Lee and Rio 2015). Other mechanisms, including usage of alternative promoters and alternative polyadenylation sites also increase the diversity of transcripts and drive both quantitative and qualitative effects (Tian and Manley 2013; de Klerk and t Hoen 2015). Indeed, alternative promoters and alternative polyadenylation sites can impact on mRNA 5'- and 3'- untranslated regions, which can have consequences on transcript stability or translation (Tian and Manley 2013; de Klerk and t Hoen 2015). Meanwhile, alternative splicing can lead to the biogenesis of nonproductive mRNAs degraded by the nonsense mediated mRNA decay pathway (Hamid and Makeyev 2014). These mechanisms can also change the gene message. Alternative promoters and alternative polyadenylation sites can impact on protein N- and C-terminal domains, respectively, and alternative splicing can impact on any protein feature (Kelemen et al. 2013; Light and Elofsson 2013; Tian and Manley 2013; de Klerk and t Hoen 2015). Therefore, these mechanisms increase the diversity of the proteome coded by a limited number of genes.

The nature (i.e., exon content) of the gene products is tightly regulated, leading different cell types to express specific sets of protein isoforms contributing to specific cellular functions. For example, the selective expression of protein isoforms plays a major role in the biological functions of epithelial and mesenchymal cells, which are two major cell types found in many tissues (Bebee et al. 2014; Mallinjoud et al. 2014; Yang et al. 2016b). Epithelial and mesenchymal cells ensure different physiological functions (epithelial cells are interconnected and non-motile cells, while mesenchymal cells are isolated and motile cells) and the epithelial-mesenchymal transition has been shown to contribute to metastasis formation during tumor progression (Bebee et al. 2014; Yang et al. 2016b).

Several splicing factors, including ESRP1, ESRP2, RBM47 and RBFOX2, control the exon inclusion rate in an epithelial- or mesenchymal-specific manner leading to the production of protein isoforms driving biological processes like cell polarity, adhesion or motility (Venables et al. 2013; Bebee et al. 2014; Mallinjoud et al. 2014; Vanharanta et al. 2014; Yang et al. 2016b).

Alternative splicing plays a major role in several pathological situations as massive splicing variations are observed in many diseases (Cieply and Carstens 2015; Daguenet et al. 2015; Sebestyen et al. 2016). However, the analysis of the cellular functions driven by specific splicing-derived protein isoforms is a major challenge for two main reasons. First, dozens of splicing variants of any one gene are often observed to be differentially expressed when comparing two biological situations. There is therefore a problem of resource prioritization for the massive task of splice isoform functional characterization. In this context, the selection of specific splicing variants for further functional analyses is often biased and based on the gene functions described in the literature, which puts the focus on well-characterized genes while overlooking the poorly characterized ones. In addition, the protein features impacted by alternative splicing are currently mostly analyzed manually in a time consuming process. The second challenge for the splicing field is the concept of 'meta-analysis' of splicing data; the functional output resulting from splicing variant mis-regulation is currently analyzed on a gene-by-gene basis without considering the global impact of co-regulated splice variants. It is expected that identifying common protein features affected by splicing variations will allow a better understanding of the contribution of alternative splicing variations in cellular phenotypes.

In order to address these concerns, we developed, and have made available on the web, a computational ontology-based approach named "Exon Ontology" (EO), that is conceptually similar to the "Gene Ontology" approach but focuses on exon level functional features instead of gene level annotations. This strategy allowed us to characterize individual and co-regulated protein features impacted by alternative splicing of exons that are differentially spliced between epithelial and mesenchymal cells.

Results

Principle of Exon Ontology: From Exon Ontology tree to scoring and statistical analysis

Large-scale RNA sequencing technologies allow to characterize the expression level of cellular transcripts as well as their exon content. Computational analyses based on Gene Ontology (GO), which relies on gene functional annotations (or GO terms), allow the prediction of the biological processes (enriched GO terms) that are likely to be impacted by changes in gene expression level (Figure 1A). We developed "Exon Ontology" to identify protein domains and features that are impacted by alternative splicing variations with the aim of predicting the contribution of alternative splicing to cellular phenotypes (Figure 1A). For this purpose, we defined Exon Ontology (EO) terms from existing ontologies and databases including Gene Ontology, Sequence Ontology, Protein Modification Ontology and InterPro (Montecchi-Palazzi et al. 2008; Mungall et al. 2011; Gene Ontology 2015; Mitchell et al. 2015) organized in an ontology tree (Figures 1B and 1C).

This Ontology tree is based on 8 major protein features that can be affected by alternative splicing (Figure 1C). This includes protein domains with catalytic, binding, receptor, and transporter activities and protein regions containing protein sub-cellular localization signals, structural features and experimentally validated post-translational modifications (PTMs). Each class of protein features was next divided into categories based on existing ontological trees. For example, the "localization" class was divided into seven categories using the ontology tree defined by the "Sequence Ontology" resource (Mungall et al. 2011) (SO, Figure 1C). Categories corresponding to the "catalytic" class were extracted from InterPro and "Gene Ontology" (Gene Ontology 2015; Mitchell et al. 2015). A total of 5,312 Exon Ontology terms (EO terms) was used to generate the Exon Ontology tree (Figure 1C and Supplementary Table S2).

Meanwhile, protein annotations retrieved from reference tools and databases were mapped to the genomic exons defined in the FasterDB genome annotation database that we previously developed (Mallinjoud et al. 2014) (Figure 1B). So doing, FasterDB genomic exons were associated with one or several EO terms and a web interface was developed in order to easily retrieve the EO terms associated with genomic exons (Figure 1B). A large proportion of the 190,617 coding exons defined in FasterDB was associated with "Structure"-, "PTM" (post-translational modification)-, "Binding"-, "Localization"-, and/or "Catalytic"-associated terms (Figure 1D). All the information, in particular the association between genomic exons and EO terms was stored in a relational MySQL database, making very easy to associate one single exon or a list of exons with EO terms, as it will be illustrated throughout the manuscript.

Predicting the impact of alternative splicing on biological processes does not only require to describe the potential protein feature(s) associated with one or several exons but it also requires to look for potential enrichment of specific protein features (or EO terms) within a list of coregulated exons. Such an analysis necessitates first a quantitative measurement (or score) of each EO term in a set of exons and, second, a statistical analysis by comparing the obtained scores to the scores obtained from control exons. To achieve this goal, we first established an EO score for each EO term by measuring the coverage of each EO term in a list of exons. The EO score is the number of nucleotides covered by hits of the EO term divided by the total number of nucleotides of all the exons from the tested list (Figure 2A and Materials and Methods). We also established a Z-score associated with a statistical test by comparing the calculated score obtained from a selected exon set to scores obtained by randomly built exon sets of approximately the same total size (Figure 2A and Materials and Methods).

In an attempt to decide what kind of control exons should be used, we generated three categories of exons (first, internal, and last coding exons) as we anticipated that the protein features encoded by exons may depend on their position within the gene. We therefore calculated the Z-score for EO terms in each of the three exon categories by comparing each of them to all the coding

exons defined in FasterDB. This revealed that different EO terms are enriched (positive Z-score values) in different parts of the mRNAs as illustrated in Figure 2B (and see Supplementary Table S3). In addition, when comparing annotated alternative internal coding exons (or alternatively spliced exons, ASE) to constitutive internal coding exons (CE), we also observed differential EO term enrichment and confirmed several previous findings (Figure 2C and Supplementary Table S3). For example, there was a strong enrichment for the "Intrinsically Unstructured Protein Regions" (IUPR) term in alternative exons when compared to constitutive exons (Figure 2C, a positive or negative Z-score value mean that an EO term is enriched in ASE or CE, respectively). This result supports previous reports indicating that alternative exons often code for intrinsically disordered protein regions (Romero et al. 2006; Buljan et al. 2012; Ellis et al. 2012; Weatheritt et al. 2012; Buljan et al. 2013; Colak et al. 2013). Meanwhile, CE are enriched for the "Polypeptide conserved regions" term when compared to ASE, supporting previous reports indicating that CE are often more conserved than ASE (Plass and Eyras 2006; Lev-Maor et al. 2007; Mudge et al. 2011). Several terms associated with "Localization" were enriched in CE when compared to ASE, however there was enrichment for several terms associated with "membrane" in ASE (Figure 2C, "Intramembrane Polypeptide Region" or IPR and Supplementary Table S3). This observation suggests that alternative splicing may impact on the ability of proteins to be incorporated into cellular membranes as it was reported in few cases (Stamm et al. 2005; Jones et al. 2009; Tejedor et al. 2015).

Even though we do not know yet the biological meaning of the enrichment for some protein features in different exon categories, we believe these data are important to underscore the importance of using an appropriate set of control exons, as exons from different categories code for different protein features. For example, with the aim of identifying protein features (i.e., EO terms) enriched in a selected list of coregulated exons, it might be better to compare exons corresponding to alternative promoters to the "First Coding Exons" category, or to compare splicing regulated exons to either constitutive or alternative internal coding exons.

Exon Ontology reveals specific protein features affected in exons that are differentially spliced between epithelial and mesenchymal cells.

To better predict the biological role of alternative splicing in epithelial and mesenchymal cells, we established a list of differentially spliced exons by comparing epithelial- and mesenchymal-like cells and applied the Exon Ontology approach to this list. For this purpose, we used several large-scale datasets (Supplementary Table S1) and selected exons that are differentially spliced when comparing normal mesenchymal to normal epithelial cells as well as breast cancer mesenchymal-like cells (from the Claudin-low subtype) to breast cancer epithelial-like cells (from the Luminal subtype). This established a list of 81 differentially spliced exons (Supplementary Table S4) that we initially validated by RT-PCR using total RNAs extracted from 4 normal epithelial and 4 normal mesenchymal cell types, and 4 Luminal (epithelial-like) and 4 Claudin-low (mesenchymal-like) cell types. A very good correlation was obtained when comparing RT-PCR and RNAseq exon inclusion (percent spliced in – PSI) rate variations (change in splicing/'delta PSI') (Figure 3A and Supplementary Figure S1 and Table S4). The inclusion rate of the 81 selected exons separately clustered epithelial-like (normal and cancer) and mesenchymal-like (normal and cancer) cell types, suggesting that this set of exons may drive general properties of epithelial- and mesenchymal-like cells (not shown). In addition, using the same datasets we identified 6 splicing factors whose expression differed when comparing epithelial- and mesenchymal-like cells. As already reported (Venables et al. 2013; Bebee et al. 2014; Mallinjoud et al. 2014; Vanharanta et al. 2014; Yang et al. 2016b), ESRP1, ESRP2 and RBM47 were more expressed in epithelial-like cells as confirmed by RT-qPCR and western blot analysis, while MBNL1, MBNL2 and RBFOX2 were more expressed in mesenchymal-like cells (Supplementary Figures S2A-S2C). As shown on Figure 3B (and see Supplementary Figures S2D, S3 and Table S4), ESRP1 and ESRP2 depletion in epithelial-like cells switched the splicing pattern from an epithelial- to a mesenchymal-like pattern for 36 exons, as did RBM47 depletion for 13 exons. In contrast, MBNL1 and MBNL2 depletion in mesenchymal–like cells switched the splicing pattern from a mesenchymal- to an epithelial-like pattern for 29 exons, as did RBFOX2 depletion for 37 exons (Figure 3B and

Supplementary Figures S2D, S3 and Table S4). Some redundancy was observed since, for example, most of the exons regulated by MBNL1 and MBNL2 are also regulated by RBFOX2 (Figure 3B). Most of the exons that are more included in mesenchymal-like cells are regulated by MBNL1 and MBLN2 and/or RBFOX2, while most of the exons more included in epithelial-like cells are regulated by ESRP1 and 2 and/or RBFOX2 (Figure 3C).

Applying the EO suite to the 81 selected exons (referred to below as the "Mes-Epi exons" list), we first noticed that all these exons are internal coding exons ("Mapping" in Supplementary Table S4) and encode for protein subcellular localization signals, protein-protein interacting domains, and/or phosphorylated peptides ("Exon Annotations" in Supplementary Table S4). Interestingly, several protein features are selectively impacted by alternative splicing differences between mesenchymal and epithelial cells when comparing this set of exons to either CE or ASE control exon sets.

Although the "Localization" term was not enriched in the "Mes-Epi exons" list compared to constitutive or alternative exons (CE or ASE), enrichment for the "Nuclear Localization Signal" (NLS) term was observed (Figure 4A and "Functional features" in Supplementary Table S4). This suggests that epithelial- and mesenchymal-like cells may express a similar set of proteins but with different sub-cellular localization since NLS-containing exons are differentially included in both cell types (Figure 4B). For example, the EO suite identified a putative NLS encoded by exon 15 of the *SLK* gene that produces a cytoplasmic kinase involved in cytoskeleton remodeling and cell migration (Al-Zahrani et al. 2013) ("Exon Annotations" in Supplementary Table S4 and Supplementary Figure S4A). As *SLK* exon 15 is more included in epithelial- than in mesenchymal-like cells (Figure 4B, comparing for example MDA-MB-231 to MCF-7 cells), we anticipated that SLK protein staining should be more pronounced in the nucleus of epithelial cells. As expected, immunofluorescence staining revealed a more restricted nuclear localization of SLK in MCF-7 (epithelial-like) than in MDA-MB-231 (mesenchymal-like) cells (Figure 4C). To further challenge the role of SLK exon 15 coding sequence, MCF-7 cells were transfected with oligonucleotides inducing SLK exon 15 skipping (TOSS E15)

combined with siRNA specifically targeting SLK exon 15 (siRNA E15), leading to the decrease of the SLK exon 15-containing transcripts (Supplementary Figure S5A). As predicted, the SLK protein staining in MCF-7 cells was less restricted to the nucleus in these conditions (Figure 4D).

In conclusion, EO reveals specific protein features (e.g. sub-cellular localization signals) affected by alternative splicing within a list of co-regulated exons. Getting automated computational assistance for predicting protein features impacted by alternative splicing ("Exon Annotations" in Supplementary Table S4 and Supplementary Figure S4) will speed up the functional analysis of protein isoforms.

Regulation of the AKT signaling pathway by epithelial-enriched splicing factors.

As already mentioned, the EO database contains experimentally validated post-translational modifications, including phosphorylation sites retrieved from several databases (see Material and Methods). The EO based-analysis revealed that the 81 selected exons are enriched for the "Phosphorylated residue", "O-phospho-L-serine" and "O-phospho-L-threonine" terms, but not for the "O4'-phospho-L-tyrosine" term when compared to CE or ASE (Figure 5A and "Functional features" in Supplementary Table S4). About one third (i.e., 28) of the protein segments coded by the 81 "Mes-Epi exons" contain at least one experimentally validated phosphorylation site ("PTM annotation" in Supplementary Table S4, Figure 5B). Interestingly, the identified phosphosites are often associated with other protein features like subcellular localization or protein-protein interacting domains (Supplementary Figures S4B, S4C, and S5B and see below).

As the EO web suite provides the surrounding sequences of post-translationally modified residues present in the selected exon set ("PTM annotation", Supplementary Table S4), we looked for potential phosphorylation site consensus sequences in "Mes-Epi exons" using the PhosphoSite website (http://www.phosphosite.org/homeAction.action). Remarkably, the LOGO obtained is very similar to the AKT signaling pathway consensus sequence defined as RXRXXS/T (Toker 2008)(Figure 5C). We also noticed that a large proportion of the phosphorylation sites are encoded by exons that

are more included in epithelial-like cells (Figure 5D) and are often regulated by epithelial-enriched splicing factors, in particular by ESRP1 and ESRP2 (Figures 5E and 5F).

Based on these observations, we tested whether the AKT signaling pathway was impacted by depletion of ESRP1 and ESRP2 in MCF-7 epithelial-like cells. Because the potential AKT-targeted phosphorylation sites are often within exons that are skipped upon ESRP depletion (green exons on Figure 5F), we anticipated that the AKT signaling pathway could be impaired in the absence of ESRP splicing factors. As expected, ESRP1 and ESRP2 depletion specifically decreased the AKT-dependent phosphorylation of AKT-downstream targets, including 4EBP1 and the ribosomal S6 protein, after cell treatment with the SC79 AKT-activator (Figure 5G, comparing lanes 4 and 3; Figure 5H).

To test whether the ESRP-mediated effect on the AKT signaling pathway was a consequence of splicing regulation, we focused on the *TSC2* gene that is known to play a major role in the AKT signaling pathway (Inoki et al. 2002; Cai et al. 2006; Toker 2008) and whose exon 27 is skipped upon ESRP-depletion (Figure 5F). *TSC2* exon 27 skipping was induced in MCF-7 cells using targeted oligonucleotides combined with exon 27-specific siRNAs (Supplementary Figure S5A). Strikingly, this resulted in the decrease in AKT-mediated phosphorylation of 4EBP1, as did ESRP-depletion (Figure 5I, comparing lanes 3 and 2).

In conclusion, the EO approach revealed that exons differentially spliced between epithelial- and mesenchymal-like cells code for protein segments containing phosphorylated residues (Figures 5A and 5B) and that the splicing events regulated by ESRP1 and ESRP2 play an important in the AKT signaling pathway in epithelial cells (Figures 5C-5F), as experimentally validated (Figures 5G-5I).

Interplay between autophagy and mesenchymal-enriched splicing factors.

Analyzing the protein features encoded by the “Mes-Epi exons”, we noticed that the EO score corresponding to “Structure” and “Secondary structure” terms was low when comparing to CE or ASE, while the “IUPR” (“Intrinsically Unstructured Protein Region”) score was slightly higher (Figure 6A). This is interesting to underline as intrinsically disordered protein regions play a very

important role in protein-protein interactions that are regulated by phosphorylation (Fukuchi et al. 2011; Colak et al. 2013; Oldfield and Dunker 2014; Uversky 2015). Remarkably, more than 82% of the phosphorylation sites present in the 81 exons are within IUPR and/or annotated “protein binding” regions (Figure 6B and Supplementary Table S4 and Figure S4C). In addition, these regions contain “P-rich” and “RXXK” motifs that are recognized by proteins, like GRB2 containing SH3 domains, involved in various signaling pathways (Belov and Mohammadi 2012) (Supplementary Figure S6A and Table S4). The co-occurrence of phosphorylated residues, IUPRs and/or protein binding motifs in the protein segments coded by the 81 “Mes-Epi exons” suggested that alternative splicing of these exons may affect protein-protein interaction networks.

We therefore looked within the IntAct database (http://www.ebi.ac.uk/intact) for the partners of the 81 proteins harboring differentially spliced “Mes-Epi exons”. Interestingly, these partners are involved in biological processes relying on “non-membrane bounded organelles” and “vesicles”, “autophagy vacuole” and “exocytosis” (Supplementary Figure S6B and S6C) and several genes bearing exons regulated by mesenchymal-enriched splicing factors interact with autophagic factors (Figures 6C, 6D and 6E). This includes for example the *KIAA0226* (or *Rubicon*) gene that codes for a major autophagy inhibitor interacting with beclin 1 (BECN1) and *WDFY3* (or *Alfy*) gene that codes for an important adaptor protein for selective autophagy interacting with LC3, GABARAP and p62 proteins (Matsunaga et al. 2009; Isakson et al. 2013; Lamb et al. 2013; Baixauli et al. 2014; Wild et al. 2014; Khaminets et al. 2016; Ktistakis and Tooze 2016).

Based on these observations, we investigated the role of mesenchymal-enriched splicing factors in autophagy, a process involved in the degradation and recycling of cellular components, in particular under cellular starvation (Matsunaga et al. 2009; Isakson et al. 2013; Lamb et al. 2013; Baixauli et al. 2014; Wild et al. 2014; Khaminets et al. 2016; Ktistakis and Tooze 2016). Autophagy is a dynamic process of intracellular bulk degradation in which cytosolic proteins and organelles are sequestered into double membrane vesicles called autophagosomes, to be fused with lysosomes for degradation and recycling. Autophagy receptors, such as p62/SQSTM1 recognize autophagic cargo

and, via binding to small ubiquitin-like modifiers such as LC3 and GABARAPs mediate formation of autophagosomes (Matsunaga et al. 2009; Isakson et al. 2013; Lamb et al. 2013; Baixauli et al. 2014; Wild et al. 2014; Khaminets et al. 2016; Ktistakis and Tooze 2016). The effect of the depletion of mesenchymal-enriched splicing factors on autophagy was tested by western blot analysis of LC3 (whose level of lipidation can be traced by the appearance of the LC3-II form) , and of the autophagy receptor p62 (SQSTM1), that is a standard marker of cellular autophagy activity as it is degraded in the autophagosome with its cargos (Matsunaga et al. 2009; Isakson et al. 2013; Lamb et al. 2013; Baixauli et al. 2014; Wild et al. 2014; Khaminets et al. 2016; Ktistakis and Tooze 2016).

As shown on Figure 6F, depletion of MBNL1, MBNL2 and RBFOX2 (siM+R) in MDA-MB-231 cells affected both the p62 and LC3 protein expression pattern. In particular, mesenchymal-enriched splicing factor depletion enhanced the decrease of p62 stimulated by serum starvation with Earle's Balanced Salt Solution (EBSS), which is classically used to activate autophagy (left panel, compare lane 3 to lane 2, see right panel for quantification). This was not due to a decrease in p62 mRNA level (Supplementary Figure S5C). Depletion of mesenchymal-enriched splicing factors under starvation conditions also affected the LC3 protein expression pattern inducing a slight increase and decrease in the LC3-I and LC3-II form level, respectively when compared to EBSS treatment alone (Figure 6F, compare lane 3 to lane 2). This could result from LC3-II degradation along with p62, since we observed an increase in total LC3 mRNA levels (Figure 6F, right panel), which in turn may contribute to the slight LC3-I form increase. Altogether, these results show that in the absence of MBNL1, MBNL2 and RBFOX2, autophagy is stimulated as evidenced by p62 and LC3-II protein levels.

To test whether the effect of mesenchymal-enriched splicing factors was a consequence of splicing regulation, we focused on the *KIAA0226* (or Rubicon) gene that is major regulator of autophagy as described above and whose exon 14 is included in a MBNL1/2- and RBFOX2-dependent manner (Figure 6D). Skipping of KIAA0226 exon 14 was forced in MDA-MB-231 cells using targeted antisense oligonucleotides and exon-specific siRNAs (Supplementary Figure S5A). Remarkably, KIAA0226 exon 14 skipping mimicked the effect of depletion of mesenchymal-enriched splicing

factors, enhancing the p62 protein level decrease under serum starvation (Figures 6G, compare lane 3 to lane 2). A similar effect was observed by inducing the skipping of WDFY3 exon 46 that is also regulated by mesenchymal-enriched splicing factors (Figure 6D and Supplementary Figure S5D).

Interestingly, manipulation of *KIAA0226* exon 14 splicing also resulted in the decrease in the MBNL1 and MBNL2 protein levels (Figure 6H, left panel), without affecting their mRNA level (Figure 6H, right panel) and mimicked the splicing effects induced by MBNL1/2 silencing (Figure 6I). These results support a model where mesenchymal-enriched splicing factors control alternative splicing of autophagic regulators that in turn regulate MBNL1 and MBNL2 expression.

In conclusion, the EO computational approach discovered specific protein features in exons that are differentially spliced between epithelial- and mesenchymal-like cells (e.g., NLS, Figure 4). This approach also revealed common protein features encoded by co-regulated exons (e.g., phosphosites, Figure 5), which allowed to discover the signaling pathways and biological processes in which these exons are involved (e.g., AKT signaling pathway and autophagy, Figures 5 and 6). In this context, it must be underlined that a GO term analysis of the genes bearing the 81 “Mes-Epi exons” did not pinpoint these molecular pathways but instead identified “GTPase regulator activity”, “cytoskeleton”, or “protein transport” molecular pathways (Supplementary Figure S5E). Because a computational approach allowing to predict the protein features affected by alternative splicing will be useful to the research community, we created a freely available web interface (http://fasterdb.ens-lyon.fr/ExonOntology/) that, after uploading the genomic coordinates of selected exons, allows to get the information stored in the Exon Ontology database, to get potentially enriched protein features, and to retrieve relevant protein-protein networks (Supplementary Figure S7).

Discussion

Alternative splicing is the main mechanism increasing the diversity of the proteome coded by a limited number of genes (Nilsen and Graveley 2010; Kelemen et al. 2013; Light and Elofsson 2013; Tian and Manley 2013; de Klerk and t Hoen 2015; Lee and Rio 2015). Transcriptomic genome-wide analyses from physiological or pathological biological samples generally uncover massive variations of alternative splicing (Cieply and Carstens 2015; Daguenet et al. 2015; Sebestyen et al. 2016). Being able to routinely measure massive splicing variations, the main challenge is now to determine how these splicing variations drive physiological- and pathological-associated cellular phenotypes. To address this challenging task, we developed the "Exon Ontology" approach that, while conceptually similar to the "Gene Ontology" approach, specifically focuses on exon functional features instead of gene functional annotations. To do so, we methodically associated human genomic exons to encoded protein features (named EO terms) using an Ontology tree approach and using already defined ontology terms based on reference resources (Montecchi-Palazzi et al. 2008; Mungall et al. 2011; Gene Ontology 2015; Mitchell et al. 2015). We also implemented an EO Z-score allowing to measure a potential EO term enrichment within a list of selected exons compared to the appropriate set of control exons (Figures 1 and 2). As we showed that different exon categories (e.g. first, internal or last coding exons, constitutive and alternative exons) are differentially enriched for specific EO terms (Figure 2), we want to stress here that it is important to compare a list of selected exons to the appropriate control list (e.g., first coding exons must be compared to the control list made with all first coding exons). The EO web suite provides support for this specific step (Supplementary Figures S7C and S7H).

One of the main powers of EO is that it allows an automatic retrieval of protein features coded by selected exons. Applying EO to a list of exons differentially spliced between epithelial- and mesenchymal-like cells allowed to uncover several exons coding for nuclear localization signals (Figure 4, Supplementary Figure S4A). A dedicated table ("Exon Annotations" in Supplementary Figure S7E) is automatically generated on the EO website and describes all the protein features

encoded by individual analyzed exons. In addition, each protein annotation can be visualized on the Fasterdb protein website (Supplementary Figure S4). This resource will therefore speed up the characterization of biological consequences resulting from splicing variations.

Another advantage of the EO approach is its ability to identify common protein features affected by coregulated exons. This allowed us to uncover a substantial number of exons differentially spliced between epithelial- and mesenchymal-like cells that code for experimentally validated phospho-site-containing protein domains (Figure 5A and "PTM annotation" in Supplementary Table 4). Looking for potential phospho-site consensus sequences allowed us to demonstrate for the first time a role of ESRP-regulated splicing events in the AKT signaling pathway in epithelial cells (Figures 5C-5I). In this setting, the Exon Ontology website generates a table containing the enrichment Z-scores for the most frequently protein features associated with the tested exons ("Functional Features" in Supplementary Figures S7H and 7I).

Finally, the EO approach uncovers protein features co-occurring within alternative exons. For example, we observed that many phospho-sites are embedded within domains involved in protein-protein interactions (e.g. protein binding motifs, intrinsically disordered regions, Figure 6B). This observation suggests that "Mes-Epi exons" exons code for protein segments playing a role in the regulation of protein interaction networks. Integrating alternative splicing and interactome datasets allowed the identification of biological processes impacted by alternative splicing as we uncovered an intricate relationship between autophagy and alternative splicing: splicing factors and alternative splicing events impact on autophagy (Figures 6F and 6G) and autophagic regulators impact on splicing factor expression and splicing decisions (Figures 6H and 6I). Therefore, the EO web resource provides the list of proteins interacting with the products of the genes bearing tested alternative exons ("Protein-Protein network", Supplementary Figure S7G).

In this context, we noticed that some splicing-regulated genes share the same interacting partners (Figure 6E and Supplementary Figure S6C) and that in such a case, the regulated exons encode for similar protein sequences (Supplementary Figure S6D). For example, the *WDFY3* and

PLOD2 genes, whose products both interact with ATG5 (Figure 6E), contain alternative exons that share strong sequence similarity (Supplementary Figure S6D). The same is true for alternative exons 11 and 7 of *EXOC1* and EXOC7 genes, respectively, whose protein products interact with EXOC4 (Figure 6E and Supplementary Figure S6D). These observations support a model where alternative exons play a role in the competition in protein interaction since EXOC1 exon 11 and EXOC7 exon 7 are regulated in an opposite manner: EXOC1 exon 11 is more included in epithelial cells and is repressed by mesenchymal splicing factors, while EXOC7 exon 7 is more included in mesenchymal cells and is positively regulated by mesenchymal splicing factors (Figure 6D). Therefore, although further experiments are needed, comparing the protein sequences encoded by coregulated exons or exons that are inversely regulated could help to identify important functional amino-acid residues.

Combined with the effort of the research community to characterize alternative splicing-dependent protein interaction networks (Corominas et al. 2014; Raj et al. 2014; Li et al. 2015; Tseng et al. 2015; Will and Helms 2016; Yang et al. 2016a) and with web services allowing to associate splicing events to protein feature annotation (Li et al. 2014; Rodriguez et al. 2015; Mall et al. 2016), the EO website will provide computational support toward the aim of developing systems-biological approaches for predicting the biological consequences resulting from splicing variations.

Materials and Methods

Ontology tree: Ontological terms have been manually selected from the Sequence Ontology – SO (version 1.45 25:08:2014), the Protein Modification Ontology – PSI-MOD (version 1.013.0 30:05:2014), and the InterPro tree and its GO mapping (version 46.0). The original ontological trees have been linked to seven main classes of protein features.

Annotations: Annotations have been derived from reference tools and databases, including interproscan (version 5.3-46.0), TMHMM (version 2.0c), IntAct (may 2015), Uniprot (oct. 2014), dbO-GAP (mar. 2014), hUbiquitome (mar. 2014), PhosphoSitePlus (apr. 2014), dbPTM3 (may 2013), PhosphoELM (may 2013), ProteomeScout (mar. 2014), D2P2 (dec. 2014). Localization motifs have been identified using a custom PERL (version 5.10.1) script based on regular expressions. These annotations have been mapped at the exon level using our splicing database FasterDB, and stored in a MySQL database (version 14.14 distribution 5.1.73).

EO score, Z-score, and FDR: For a given exon and a given feature, the EO score is computed by dividing the feature size (in nucleotides) by the exon size (in kilo-nucleotides). Only the coding part of the exon is considered. When the feature only partially overlaps the exon, only this overlapping region is considered. The Z-score is based on the comparison of an EO score of interest (for a selected set of sequences) with the distribution of 1,000 EO scores obtained with sequence sets of approximately the same size that are randomly generated from a control sequence sets (for instance from all first coding exons). This is only done for the EO terms that are annotated with at least 4% of the human exons (91 EO terms, see Supplementary Table S2). The EO score distributions have been generated off-line for sequence sets of varying sizes (from 100 nucleotides to 32 kilo-nucleotides). The EO scores are log-normally distributed so the log of the EO scores are used to compute the Z-scores. The FDR is computed using the Benjamini and Hochberg strategy.

Web interface: The web interface is written in PHP and Javascript. It also relies on a set of PERL (version 5.20.2) script to interact with the MySQL database (version 14.14 distribution 5.5.49). The web server is run by Apache (version 2.4.10) on a Debian machine (version 8.5).

Cell culture, treatment and transfection: Cell culture of standard MCF-7 and MDA-MB-231 cells, as well as transient transfection assays were performed essentially as described previously (Dardenne et al., 2012; Samaan et al., 2014). Sequences of siRNAs and TOSS are provided in Supplementary Table S1. AKT activation experiments were performed as follows: 24 hours after siRNA transfection, cells were first deprived in serum-free medium (Earle's Balanced Salts with Sodium medium, EBSS, Sigma E3024 and E2888) for 16 hours and then reactivated in medium containing 5µg/ml of SC79 (pan-AKT activator by phosphorylation, S7863 – Selleckchem) for 1 hour.

RNA analysis: RNA extraction, RT-PCR and RT-qPCR were described previously (Dardenne et al., 2012; Samaan et al., 2014). qPCR data were normalized with NUC18s gene as a control. Statistical analyses on means were performed using Student's Tests (unilateral, paired, $p<0.05$). Primer sequences for PCR and qPCR are provided in Supplementary Table S1.

Western blot analysis: Total cell extracts were lysed in "NP40 buffer" (Tris-HCl 50mM final, NaCl 400mM final, EDTA 5mM final, IGEPAL 1% final, SDS 0.2% final) complemented with protease and phosphatase inhibitors (11836145001 and 04906837001, Roche) and then incubated on ice for 30min. Extracts were then sonicated for 10 min (10 cycles, 30" on / 30" off). Protein concentrations from total cell extracts were determined using Pierce BCA Protein Assay Kit (23225, Thermo Scientific). Total cell extracts were run on 4-12% Bis-Tris gels (Invitrogen) and transferred on nitrocellulose membranes (IB301001 iBlot Gel Transfer Stacks Nitrocellulose, Invitrogen). Membranes were washed in TBST (Tris-Base 20mM final, NaCl 130mM final, 0.1% Tween 20, pH adjusted to 7.6 with 37% HCl) and blocked in 5% (w/v) dry non-fat milk or 5% (w/v) BSA (A7030, Sigma) for primary phospho-antibodies. Membranes were then incubated with primary antibodies (overnight, 4°C) and washed before to be incubated with secondary HRP-conjugated antibodies for 1 hour. Primary and secondary antibodies are listed in Supplementary Table S1. Image acquisitions were performed using ChemiDoc Touch Imaging System (Biorad) and quantification was performed using Image Lab software (v.5.2.1, Biorad) and normalized with H3 or total non-phospho-protein. Statistical analyses on means were made using Student's Tests (unilateral, paired, $p<0.05$).

Immunofluorescence: Cells were fixed in 4% paraformaldehyde for 20 min. After 3 washes in 1x PBS, cells were permeabilized in 0.2% Triton X-100 for 30 min and left for 1 h in blocking solution (1x PBS, 15% serum, 0.1% Triton X-100). Slides were then incubated in blocking solution containing rabbit anti-SLK (ab65113, Abcam, 1/100) primary antibody (overnight, 4°C). After 3 washes in 1x PBS, slides were incubated 2 hours in blocking solution containing FITC-conjugated anti-rabbit IgG (Sigma, 1/2.000) and nuclei were stained with DAPI (10nM final, 10min).

Acknowledgments

This work has been funded by Fondation ARC (Programme Labellisé Fondation ARC 2014, PGA120140200853), INCa (2014-154), Inserm "Plan Cancer 2009-2013", AFM-Téléthon, and LNCC. Doctoral fellowships from Région Rhône-Alpes (C.B.) and post-doctoral fellowships from LNCC (M.A.M.G) and Fondation ARC (F.Z.C., F.O.D. and L.C.T.). The authors are grateful for helpful discussions on statistical analysis with Claire Burny.

Figure legends

Figure 1

A. Genome-wide transcriptomic analyses allow identification of genes whose expression level is modified when comparing two experimental conditions. Looking for the enrichment of Gene Ontology (GO) terms associated with these genes allows prediction of the biological processes and cellular activities that are likely to be impacted by gene expression level modifications. Exon Ontology (EO) aims at identifying proteins features associated with changes of exon content owing to alternative splicing regulation, which may contribute to cellular phenotypes. Both GO and EO predictions can next be addressed by dedicated experimental approaches.

B. The Exon Ontology workflow is based on ontological terms (EO terms) that were derived from existing ontologies and databases (e.g. GO, Sequence ontology, PSI-MOD, and InterPro,). Protein features were derived from reference tools and databases and were mapped to annotated genomic exons in the 'Faster DB' database. Genomic exons can thus be associated with one or several EO terms (EO annotations at the exon level). A computational suite (Exon Ontology) then calculates a dedicated EO term score and looks for potential EO term enrichment by statistical analysis.

C. Protein features and domains have been assigned Exon Ontology (EO) terms based on existing ontologies and databases as described in panel B. The EO terms were organized in an Exon Ontology tree based on height classes of protein features (e.g., catalytic, binding). Each class was divided in categories and contains a more or less large number of associated terms. For example, the "Localization" class was divided into "Nuclear Localization Signal" (NLS), "Nuclear Export Signal" (NES), "Mitochondrial Targeting Signal" (MTS), "Peroxisomal Targeting Signal" (PTS), "Endoplasmic Reticulum Signal Sequence" (ERSS), "Endosomal Localization Signal" (EL), and "Signal Peptide" (SP) categories based on the "Sequence Ontology" resource.

D. Pie chart showing the distribution of functional annotations of human coding exons; more than 170,000 human coding exons are associated with at least one Exon Ontology term. The numbers represent the number of exons associated with the main classes of the Exon Ontology terms.

Figure 2

A. Looking for enrichment of protein features (EO terms) encoded by a set of exons requires a quantitative measurement. The 'EO score' of each EO term associated with exons from a test-list is calculated by dividing the number of nucleotides covered by each EO term (size(h)) divided by the total number of nucleotides of each tested exons (size(e)). The enrichment Z-score and statistical significance are calculated by comparing the calculated EO score as described above to the scores obtained from a large number of sets of control exons having approximately the same size as the

test-list. This analysis is currently available on the Exon Ontology web site for the 91 EO terms that are most frequently associated with exons (listed in Supplementary Table S2).

B. Several Exon Ontology terms were enriched at either end of their mRNAs (first, internal and last coding exons). The x-axis corresponds to the Z-scores obtained by comparing one category of exons to "all" exons. Positive Z-score values (red numbers and boxes) indicate EO term enrichment in the corresponding exon category. IUPR ("Intrinsically Unstructured Polypeptide Region"); Acetylated ("Acetylated-residues"). *P-value<0.05;***P-value<0.005.

C. Constitutive and alternative coding exons are enriched for different EO terms. The y-axis corresponds to the Z-scores obtained by comparing alternative to constitutive exons. Positive Z-scores indicate enrichment of the corresponding EO term in alternative exons, while negative values indicate enrichment in constitutive exons. IUPR ("Intrinsically unstructured polypeptide region"); IPR ("Intramembrane Protein Region"). *P-value<0.05;***P-value<0.005.

Figure 3

A. Comparison of the exon percent splicing inclusion (psi) rate variations (deltaPSI) of 81 exons as measured by RT-PCR (x-axis) and by RNAseq (y-axis) of differentially spliced exons between normal fibroblast (Fibro) and normal epithelial (Epi) cells (left panel) and breast cancer Claudin Low (mesenchymal-like) versus Luminal (epithelial-like) cells (right panel).

B. Number of exons regulated by MBNL1&2 and RBFOX2 in MDA-MB-231 cells or regulated by ESRP1&2 and RBM47 in MCF-7 cells. Significant collusion was observed amongst mesenchymal and epithelial splicing factors, respectively.

C. Number of exons more (red circles) or less (green circles) included in mesenchymal-like compared to epithelial-like cells and regulated by MBNL1&2, RBFOX2 (right) and ESRP1&2 and/or RBM47 (left).

Figure 4

A. Exons differentially spliced between epithelial- and mesenchymal-like cells ("Mes-Epi exons") code for protein segments that are enriched in "NLS" (Nuclear Localization Signal) term when compared to constitutive (CE) or alternative (ASE) exons. *p-value<0.05.

B. RT-PCR performed with total RNAs obtained from 4 normal epithelial (1=HEPic, 2=HPAEPic, 3=HMEC, 4=AG01134) and 4 normal mesenchymal (5=HMF, 6=HCFaa, 7=AG0449, 8=AG0450) cell lines and the breast cancer MCF-7 and MDA-MB-231 cell lines. The selected genes correspond to genes bearing alternative exons coding for protein segments containing nuclear localization signal (NLS). Red and green rectangle correspond to alternative exons that are more and less included, respectively in mesenchymal- compared to epithelial-like cells.

C. SLK exon 15 that encodes for a NLS is more included in MCF-7 than in MDA-MB-231 cells (see panel B). Immunofluorescence of SLK protein indicates that SLK is more restricted to the nucleus in MCF-7 (epithelial-like) than in MDA-MB-231 (mesenchymal-like) cells.

D. Depletion of SLK transcripts that contain exon 15 (TOSS E15 + siRNA E15) leads to a more diffuse staining within transfected MCF-7 cells compared to control (CTRL) cells.

Figure 5

A. Exons differentially spliced between epithelial- and mesenchymal-like cells ("Mes-Epi exons") encode for protein segments that are enriched, when compared to constitutive (CE) or alternative (ASE) exons, for "Phosphorylated residue" (Phospho-sites), "O-phospho-L-serine" (pSerine), "O-phospho-L-threonine" (pThreonine) terms but not for the "O4'-phospho-L-tyrosine" (pTyrosine) term. **p-value<0.005;***p-value<0.001.

B. RT-PCR performed with total RNAs obtained from 4 normal epithelial (1=HEPic, 2=HPAEPic, 3=HMEC, 4=AG01134) and 4 normal mesenchymal (5=HMF, 6=HCFaa, 7=AG0449, 8=AG0450) cell lines and the breast cancer MCF-7 and MDA-MB-231 cell lines. The selected genes correspond to genes bearing alternative exons coding for protein segments containing experimentally validated phospho-sites. Red and green rectangle correspond to alternative exons more and less included, respectively, in mesenchymal- than in epithelial-like cells.

C. Sequence "LOGO" generated from the "PhosphoSite" website using sequences surrounding experimentally validated phosphorylated residues coded by exons differentially spliced between epithelial- and mesenchymal-like cells.

D. 44 and 25 experimentally-validated phosphorylated residues are encoded within exons more included and less included, respectively in epithelial-like cells.

E. 40 and 26 experimentally-validated phosphorylated residues are encoded within exons regulated by epithelial-enriched splicing factors (ESRP1, ESRP2, and RBM47) and by mesenchymal-enriched splicing factors (MBNL1, MBNL2, and RBFOX2), respectively (left panel). 26 and 10 experimentally-validated phosphorylated serine residues are encoded within exons regulated by epithelial-enriched splicing factors (ESRP1, ESRP2, and RBM47) and by mesenchymal enriched splicing factors (MBNL1, MBN2, and RBFOX2), respectively (right panel).

F. RT-PCR performed with total RNAs extracted from the MCF-7 epithelial-like breast cancer cell line transfected with control siRNAs (1), siRNAs targeting ESRP1 and ESRP2 (2) or RBM47 (3). The selected genes correspond to genes bearing alternative exons encoding for experimentally validated phosphorylated residues. Red and green rectangles correspond to alternative exons more and less included, respectively in mesenchymal- than in epithelial-like cells.

G. Western blot analyses of the phosphorylation pattern of proteins involved downstream of the AKT signaling pathway in the MCF-7 epithelial-like cell line transfected with control siRNAs or siRNAs targeting ESRP1 and ESRP2 and treated, or not, for 1 hour with SC79 (AKT kinase activator). The pE4BP1(70) and pE4BP1(36&47) antibodies recognize phosphorylated residues on position 70, 26 and/or 47 of the E4BP1 protein. The pS6 antibody recognizes phosphorylated S6 protein. H3 (histone H3) is used as a loading control.

H. Quantification of Western blots shown in panel G. pE4BP1(70) and pE4BP1(36&47) signals were normalized by the signal obtained with antibody recognizing phosphorylated and un-phosphorylated E4BP1 protein (E4BP1). Likewise, the pS6 signal was normalized to total S6 signal (S6). **p-value<0.005.

I. Western blot analyses of the phosphorylation pattern of 4EBP1 protein in MCF-7 transfected, or not, with TOSS and siRNAs targeting TSC2 exon 27 (TOSS/siTSC2) and treated, or not, for 1 hour with SC79. H3 (histone H3) is used as a loading control.

Figure 6

A. Exons differentially spliced between epithelial- and mesenchymal-like cells ("Mes-Epi exons") code for protein segments poorly associated with "structure" and "secondary structure" terms, but that do contain unstructured (IUPR) regions.

B. 28 and 37 exons of the 81 selected exons code for protein segments containing experimentally validated phospho-sites, IUPRs and/or "Protein Binding" motifs, respectively.; 23 of them contain both phospho-sites and protein interacting motifs.

C. RT-PCR performed with total RNAs obtained from 4 normal epithelial (1=HEPic, 2=HPAEPic, 3=HMEC, 4=AG01134) and 4 normal mesenchymal (5=HMF, 6=HCFaa, 7=AG0449, 8=AG0450) cell lines and the breast cancer MCF-7 and MDA-MB-231 cell lines. The selected genes correspond to genes bearing alternative exons coding for protein interacting with autophagic factors. Red and green rectangle correspond to alternative exons more and less included, respectively, in mesenchymal- than in epithelial-like cells.

D. RT-PCR corresponding to genes with alternative exons and interacting with proteins involved in autophagy, using total RNAs obtained from mesenchymal-like MDA-MB-231 breast cancer cells transfected with control siRNAs (lane 1), siRNAs targeting MBNL1 and MBNL2 (lane 2) or RBFOX2 (lane 3). Red and green rectangles correspond to alternative exons that are more or less included, respectively in mesenchymal- than in epithelial-like cells.

E. Genes with exons regulated by mesenchymal-enriched splicing factors produce proteins interacting with proteins involved in autophagy. Red and green proteins correspond to genes with

alternative exons that are more and less included, respectively in mesenchymal- than in epithelial-like cells.

F. Western blot analyses of p62, LC3, MBNL1, MBNL2, and RBFOX2 in control (EBSS-) or serum starved (EBSS +) MDA-MB-231 cell line transfected with control siRNAs or siRNAs targeting MBNL1, MBNL2 and RBFOX2 (siM+R). H3 (histone H3) is used as a loading control. The quantification of the p62 Western blot signal and LC3 mRNA level by RT-qPCR is shown on the right. *p-value<0.05.

G. Western blot analyses of p62 and LC3 in control (EBSS-) or serum starved (EBSS +) MDA-MB-231 cell line transfected with TOSS and siRNA targeting KIAA0226 exon 14 (TOSS/siKIAA0226). H3 (histone H3) is used as a loading control.

H. MDA-MB-231 cells were transfected with TOSS and siRNA targeting KIAA0226 exon 14 (TOSS/siKIAA0226). Western blot analysis of MBNL1, MBNL2, and H3 (histone H3) used as a loading control (left panel). RT-qPCR analysis of the MBNL1 and MBNL2 mRNA levels in the same experimental conditions (right panel).

I. RT-PCR analysis using total RNAs extracted from MDA-MB-231 mesenchymal-like cells transfected as described in Panel H or transfected with siRNAs targeting MBNL1 and MBNL2 (siMBNL1&2).

References

Al-Zahrani KN, Baron KD, Sabourin LA. 2013. Ste20-like kinase SLK, at the crossroads: a matter of life and death. *Cell adhesion & migration* **7**(1): 1-10.

Baixauli F, Lopez-Otin C, Mittelbrunn M. 2014. Exosomes and autophagy: coordinated mechanisms for the maintenance of cellular fitness. *Frontiers in immunology* **5**: 403.

Bebee TW, Cieply BW, Carstens RP. 2014. Genome-wide activities of RNA binding proteins that regulate cellular changes in the epithelial to mesenchymal transition (EMT). *Advances in experimental medicine and biology* **825**: 267-302.

Belov AA, Mohammadi M. 2012. Grb2, a double-edged sword of receptor tyrosine kinase signaling. *Science signaling* **5**(249): pe49.

Buljan M, Chalancon G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, Babu MM. 2013. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Current opinion in structural biology* **23**(3): 443-450.

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular cell* **46**(6): 871-883.

Cai SL, Tee AR, Short JD, Bergeron JM, Kim J, Shen J, Guo R, Johnson CL, Kiguchi K, Walker CL. 2006. Activity of TSC2 is inhibited by AKT-mediated phosphorylation and membrane partitioning. *The Journal of cell biology* **173**(2): 279-289.

Cieply B, Carstens RP. 2015. Functional roles of alternative splicing factors in human disease. *Wiley interdisciplinary reviews RNA* **6**(3): 311-326.

Colak R, Kim T, Michaut M, Sun M, Irimia M, Bellay J, Myers CL, Blencowe BJ, Kim PM. 2013. Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS computational biology* **9**(4): e1003030.

Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, Broly M, Rodriguez M, Tam S, Trigg SA et al. 2014. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature communications* **5**: 3650.

Daguenet E, Dujardin G, Valcarcel J. 2015. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO reports* **16**(12): 1640-1655.

de Klerk E, t Hoen PA. 2015. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in genetics : TIG* **31**(3): 128-139.

Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM et al. 2012. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell* **46**(6): 884-892.

Fukuchi S, Hosoda K, Homma K, Gojobori T, Nishikawa K. 2011. Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC structural biology* **11**: 29.

Gene Ontology C. 2015. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**(Database issue): D1049-1056.

Hamid FM, Makeyev EV. 2014. Emerging functions of alternative splicing coupled with nonsense-mediated decay. *Biochemical Society transactions* **42**(4): 1168-1173.

Inoki K, Li Y, Zhu T, Wu J, Guan KL. 2002. TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling. *Nature cell biology* **4**(9): 648-657.

Isakson P, Holland P, Simonsen A. 2013. The role of ALFY in selective autophagy. *Cell death and differentiation* **20**(1): 12-20.

Jones DC, Roghanian A, Brown DP, Chang C, Allen RL, Trowsdale J, Young NT. 2009. Alternative mRNA splicing creates transcripts encoding soluble proteins from most LILR genes. *European journal of immunology* **39**(11): 3195-3206.

Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. *Gene* **514**(1): 1-30.

Khaminets A, Behl C, Dikic I. 2016. Ubiquitin-Dependent And Independent Signals In Selective Autophagy. *Trends in cell biology* **26**(1): 6-16.

Ktistakis NT, Tooze SA. 2016. Digesting the Expanding Mechanisms of Autophagy. *Trends in cell biology*.

Lamb CA, Yoshimori T, Tooze SA. 2013. The autophagosome: origins unknown, biogenesis complex. *Nature reviews Molecular cell biology* **14**(12): 759-774.

Lee Y, Rio DC. 2015. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual review of biochemistry* **84**: 291-323.

Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast G. 2007. The "alternative" choice of constitutive exons throughout evolution. *PLoS genetics* **3**(11): e203.

Li HD, Menon R, Omenn GS, Guan Y. 2014. The emerging era of genomic data integration for analyzing splice isoform function. *Trends in genetics : TIG* **30**(8): 340-347.

Li HD, Omenn GS, Guan Y. 2015. MIsoMine: a genome-scale high-resolution data portal of expression, function and networks at the splice isoform level in the mouse. *Database : the journal of biological databases and curation* **2015**: bav045.

Light S, Elofsson A. 2013. The impact of splicing on protein domain architecture. *Current opinion in structural biology* **23**(3): 451-458.

Mall T, Eckstein J, Norris D, Vora H, Freese NH, Loraine AE. 2016. ProtAnnot: an App for Integrated Genome Browser to display how alternative splicing and transcription affect proteins. *Bioinformatics*.

Mallinjoud P, Villemin JP, Mortada H, Polay Espinoza M, Desmet FO, Samaan S, Chautard E, Tranchevent LC, Auboeuf D. 2014. Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome research* **24**(3): 511-521.

Matsunaga K, Noda T, Yoshimori T. 2009. Binding Rubicon to cross the Rubicon. *Autophagy* **5**(6): 876-877.

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic acids research* **43**(Database issue): D213-221.

Montecchi-Palazzi L, Beavis R, Binz PA, Chalkley RJ, Cottrell J, Creasy D, Shofstahl J, Seymour SL, Garavelli JS. 2008. The PSI-MOD community standard for representation of protein modification data. *Nature biotechnology* **26**(8): 864-866.

Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J. 2011. The origins, evolution, and functional potential of alternative splicing in vertebrates. *Molecular biology and evolution* **28**(10): 2949-2959.

Mungall CJ, Batchelor C, Eilbeck K. 2011. Evolution of the Sequence Ontology terms and relationships. *Journal of biomedical informatics* **44**(1): 87-93.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**(7280): 457-463.

Oldfield CJ, Dunker AK. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry* **83**: 553-584.

Plass M, Eyras E. 2006. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. *BMC evolutionary biology* **6**: 50.

Raj B, Irimia M, Braunschweig U, Sterne-Weiler T, O'Hanlon D, Lin ZY, Chen GI, Easton LE, Ule J, Gingras AC et al. 2014. A global regulatory mechanism for activating an exon network required for neurogenesis. *Molecular cell* **56**(1): 90-103.

Rodriguez JM, Carro A, Valencia A, Tress ML. 2015. APPRIS WebServer and WebServices. *Nucleic acids research* **43**(W1): W455-459.

Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z et al. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences of the United States of America* **103**(22): 8390-8395.

Sebestyen E, Singh B, Minana B, Pages A, Mateo F, Pujana MA, Valcarcel J, Eyras E. 2016. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome research* **26**(6): 732-744.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* **344**: 1-20.

Tejedor JR, Papasaikas P, Valcarcel J. 2015. Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis. *Molecular cell* **57**(1): 23-38.

Tian B, Manley JL. 2013. Alternative cleavage and polyadenylation: the long and short of it. *Trends in biochemical sciences* **38**(6): 312-320.

Toker A. 2008. mTOR and Akt signaling in cancer: SGK cycles in. *Molecular cell* **31**(1): 6-8.

Tseng YT, Li W, Chen CH, Zhang S, Chen JJ, Zhou X, Liu CC. 2015. IIIDB: a database for isoform-isoform interactions and isoform network modules. *BMC genomics* **16 Suppl 2**: S10.

Uversky VN. 2015. Functional roles of transiently and intrinsically disordered regions within proteins. *The FEBS journal* **282**(7): 1182-1189.

Vanharanta S, Marney CB, Shu W, Valiente M, Zou Y, Mele A, Darnell RB, Massague J. 2014. Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. *eLife* **3**.

Venables JP, Brosseau JP, Gadea G, Klinck R, Prinos P, Beaulieu JF, Lapointe E, Durand M, Thibault P, Tremblay K et al. 2013. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Molecular and cellular biology* **33**(2): 396-405.

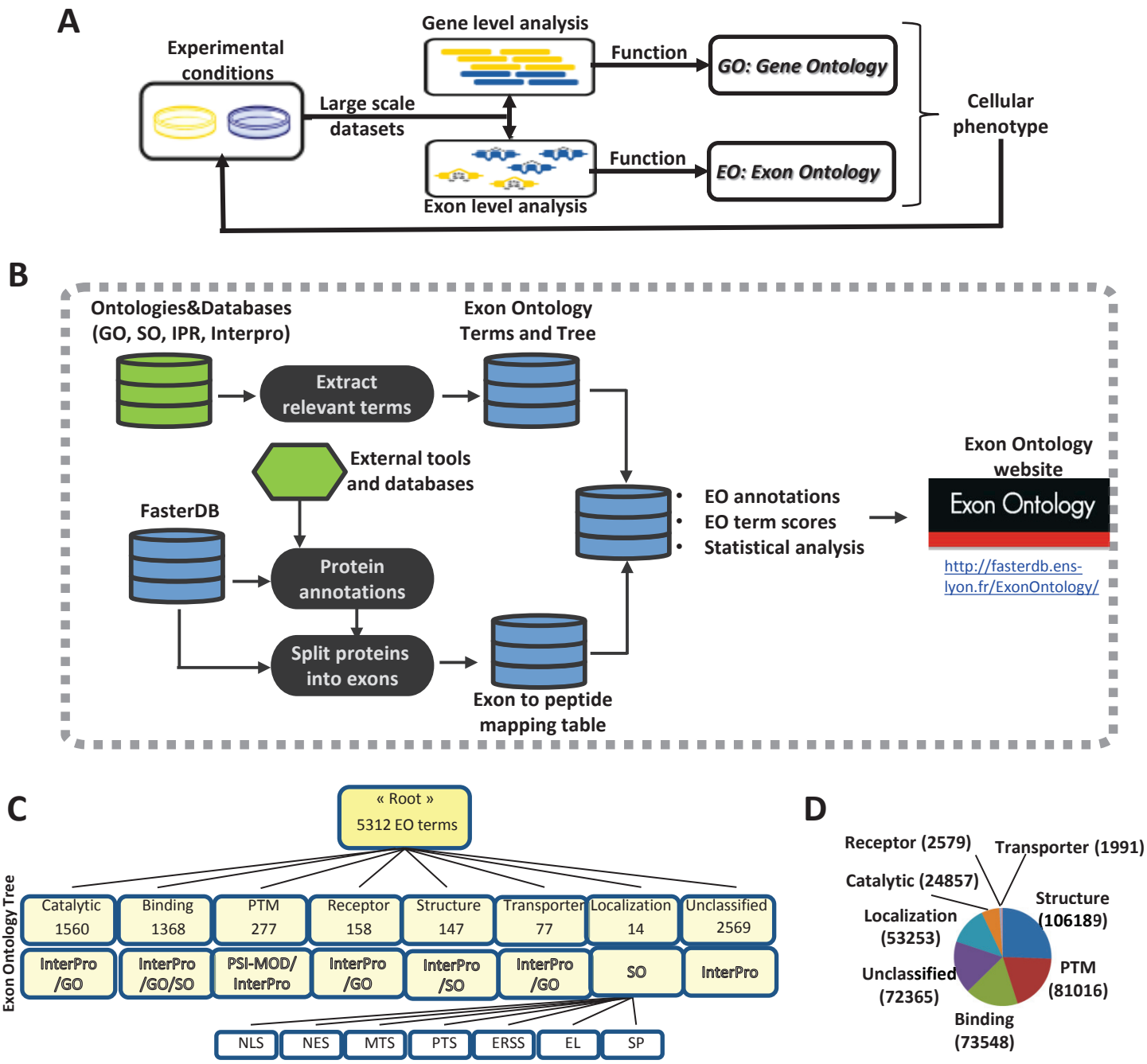
Weatheritt RJ, Davey NE, Gibson TJ. 2012. Linear motifs confer functional diversity onto splice variants. *Nucleic acids research* **40**(15): 7123-7131.

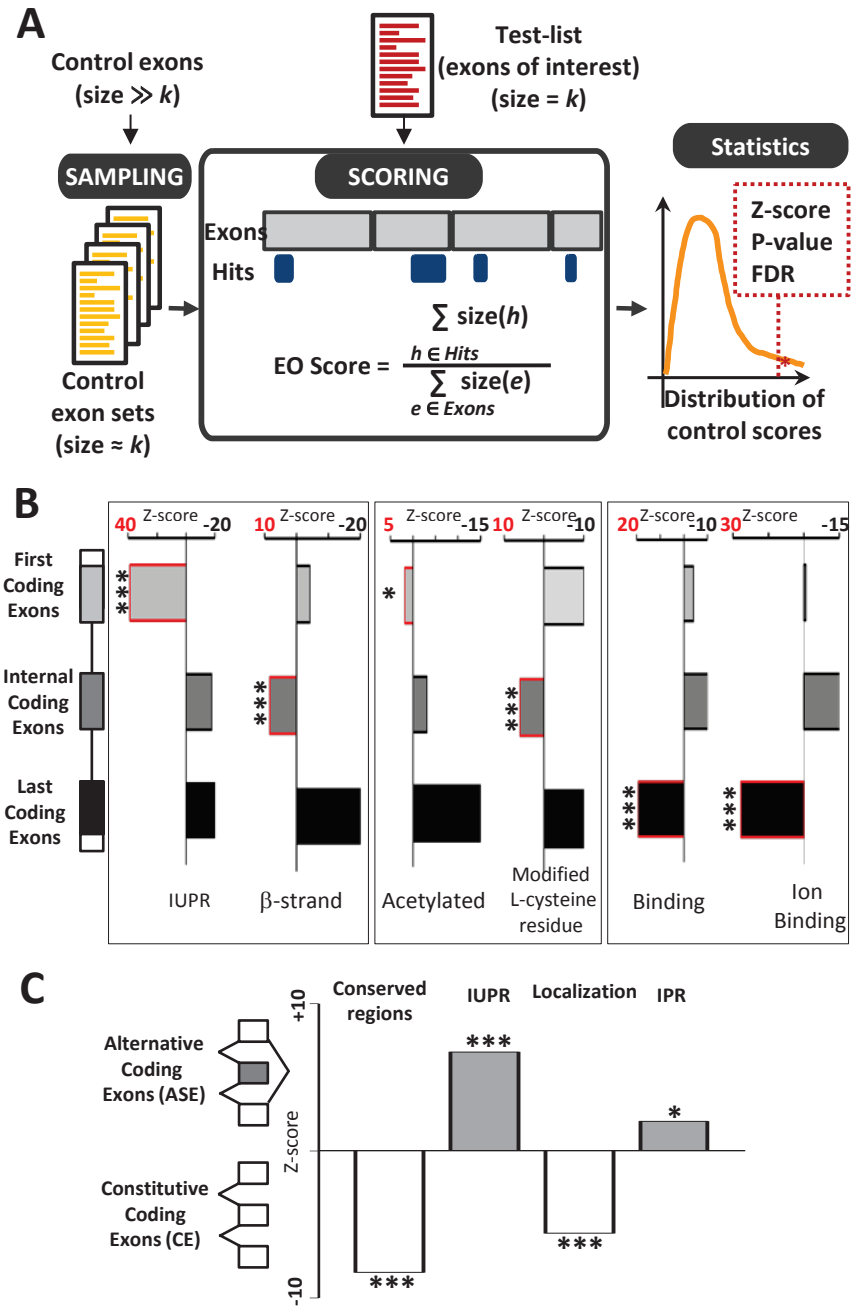
Wild P, McEwan DG, Dikic I. 2014. The LC3 interactome at a glance. *Journal of cell science* **127**(Pt 1): 3-9.

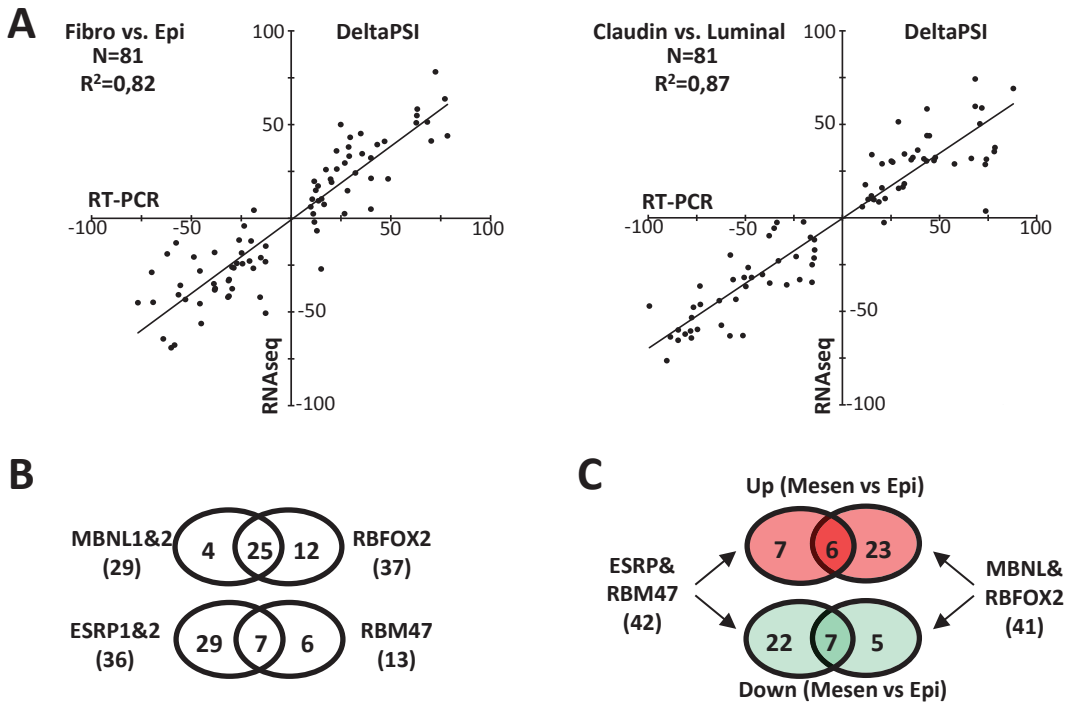
Will T, Helms V. 2016. PPIXpress: construction of condition-specific protein interaction networks based on transcript expression. *Bioinformatics* **32**(4): 571-578.

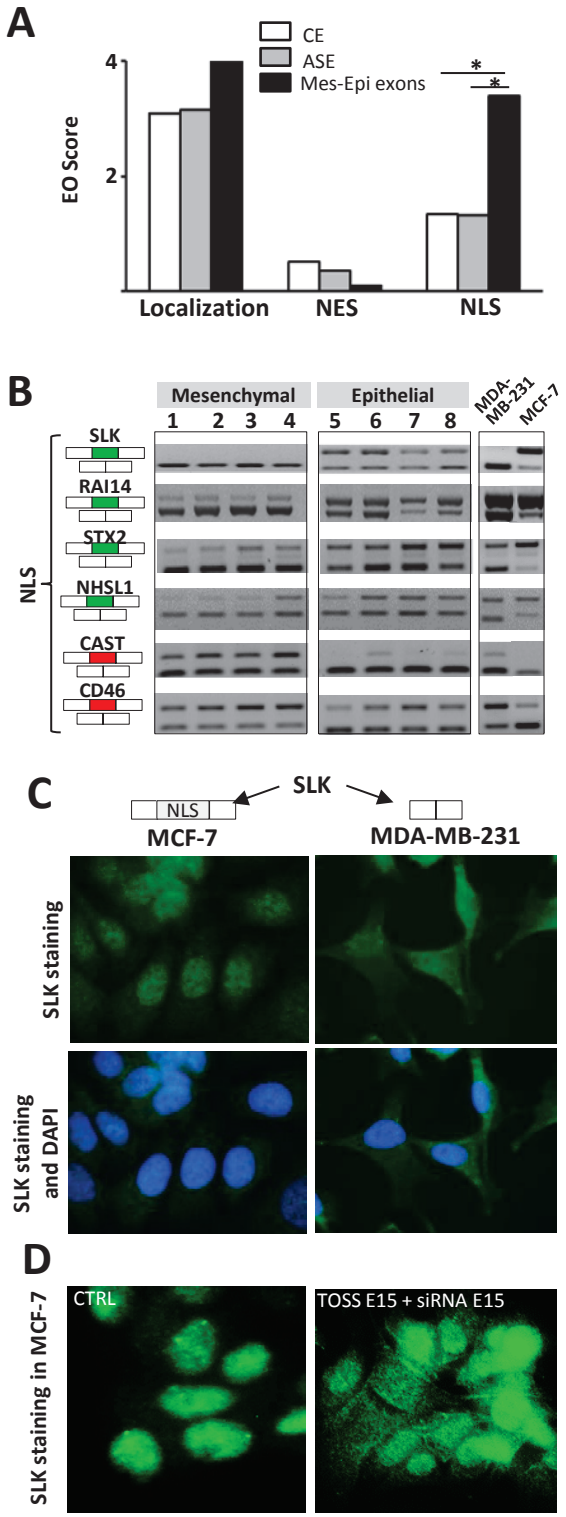
Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR et al. 2016a. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**(4): 805-817.

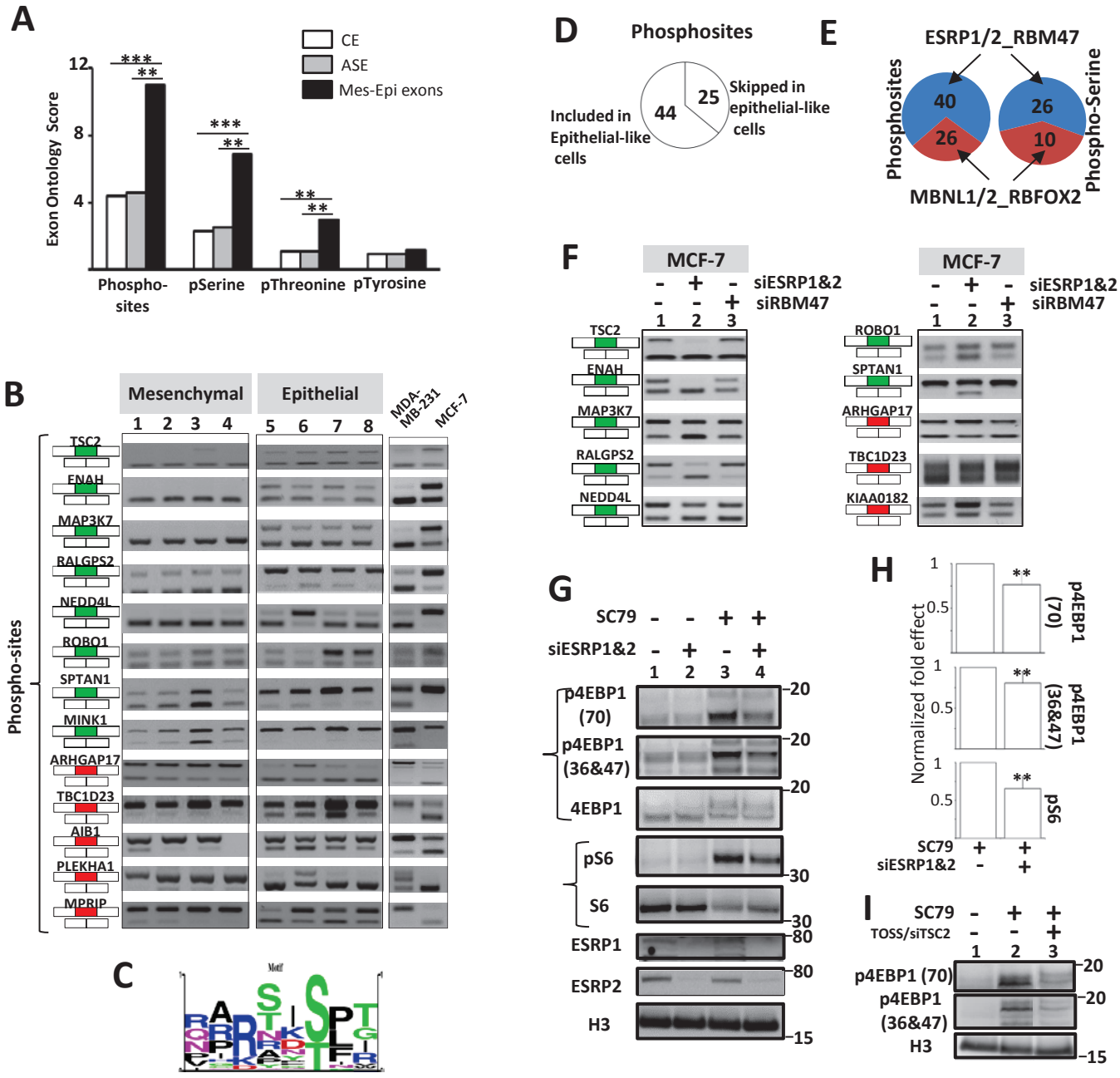
Yang Y, Park JW, Bebee TW, Warzecha CC, Guo Y, Shang X, Xing Y, Carstens RP. 2016b. Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. *Molecular and cellular biology* **36**(11): 1704-1719.

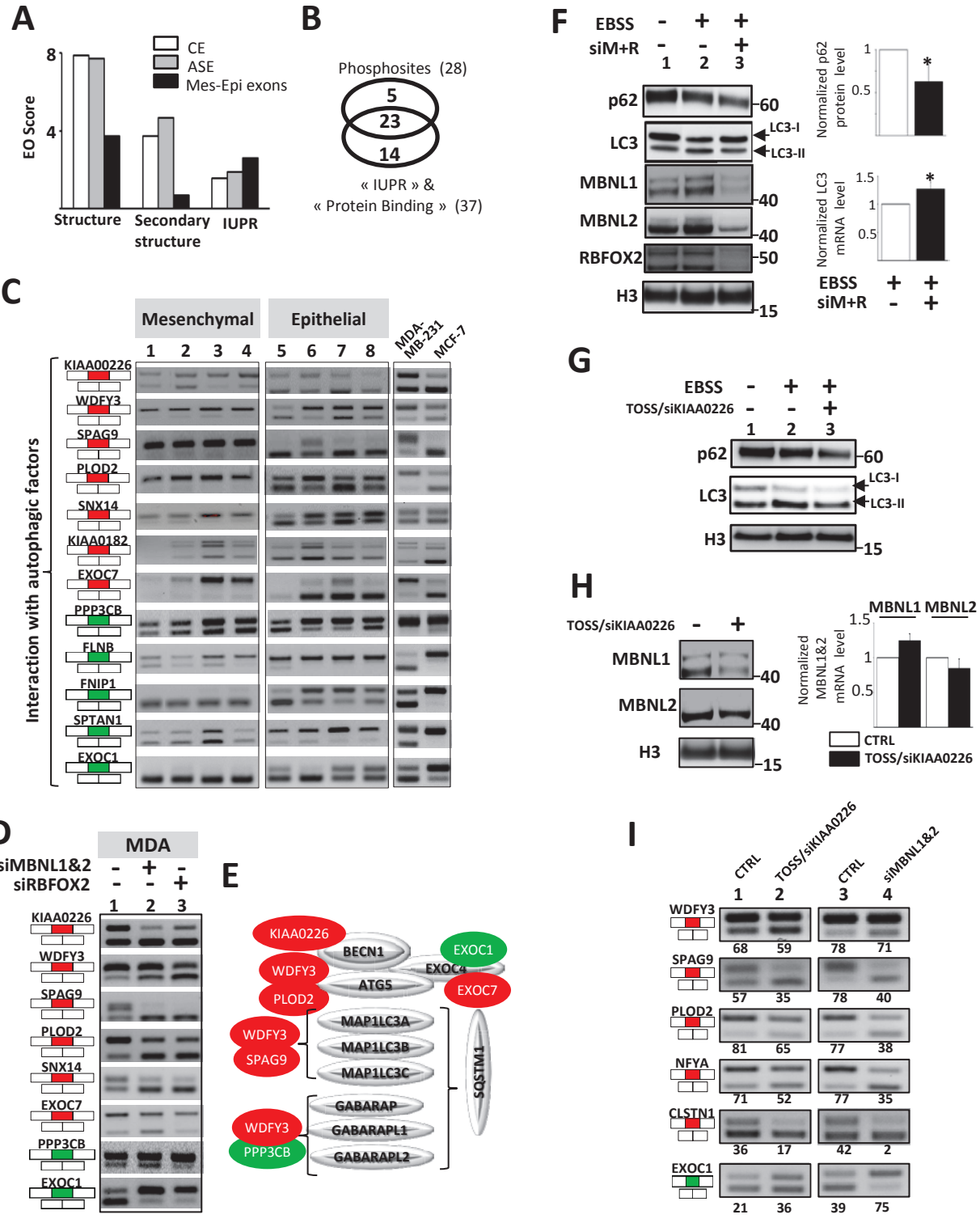

A
Experimental conditions
Large scale datasets
Gene level analysis
Exon level analysis
Function
GO: Gene Ontology
Function
EO: Exon Ontology
Cellular phenotype
B
Ontologies&Databases (GO, SO, IPR, Interpro)
Extract relevant terms
Exon Ontology Terms and Tree
External tools and databases
FasterDB
Protein annotations
Split proteins into exons
Exon to peptide mapping table
EO annotations
EO term scores
Statistical analysis
Exon Ontology website
Exon Ontology
http://fasterdb.ens-lyon.fr/ExonOntology/
C
« Root »
5312 EO terms
Exon Ontology Tree
Catalytic 1560
Binding 1368
PTM 277
Receptor 158
Structure 147
Transporter 77
Localization 14
Unclassified 2569
InterPro /GO
InterPro /GO/SO
PSI-MOD/ InterPro
InterPro /GO
InterPro /SO
InterPro /GO
SO
InterPro
NLS
NES
MTS
PTS
ERSS
EL
SP
D
Receptor (2579)
Transporter (1991)
Catalytic (24857)
Structure (106189)
Localization (53253)
Unclassified (72365)
PTM (81016)
Binding (73548)


A
Control exons
(size ≫ k)
Test-list
(exons of interest)
(size = k)
SAMPLING
SCORING
Statistics
Exons
Hits
∑ size(h)
EO Score = h ∈ Hits / ∑ size(e) e ∈ Exons
Z-score
P-value
FDR
Control
exon sets
(size ≈ k)
Distribution of
control scores
B
First
Coding
Exons
Internal
Coding
Exons
Last
Coding
Exons
40 Z-score -20
10 Z-score -20
5 Z-score -15
10 Z-score -10
20 Z-score -10
30 Z-score -15
IUPR
β-strand
Acetylated
Modified
L-cysteine
residue
Binding
Ion
Binding
C
Alternative
Coding
Exons (ASE)
Constitutive
Coding
Exons (CE)
+10
Z-score
-10
Conserved
regions
IUPR
Localization
IPR


A
Fibro vs. Epi
N=81
R²=0,82
DeltaPSI
RT-PCR
RNAseq
100
50
-50
-100
Claudin vs. Luminal
N=81
R²=0,87
DeltaPSI
RT-PCR
RNAseq
B
MBNL1&2
(29)
4
25
12
RBFOX2
(37)
ESRP1&2
(36)
29
7
6
RBM47
(13)
C
Up (Mesen vs Epi)
ESRP&
RBM47
(42)
7
6
23
MBNL&
RBFOX2
(41)
22
7
5
Down (Mesen vs Epi)


A
CE
ASE
Mes-Epi exons
EO Score
4
2
*
*
Localization
NES
NLS
B
Mesenchymal
Epithelial
MDA-MB-231
MCF-7
1 2 3 4
5 6 7 8
NLS
SLK
RAI14
STX2
NHSL1
CAST
CD46
C
NLS
SLK
MCF-7
MDA-MB-231
SLK staining
SLK staining and DAPI
D
SLK staining in MCF-7
CTRL
TOSS E15 + siRNA E15


A
CE
ASE
Mes-Epi exons
Exon Ontology Score
12
8
4
***
**
Phospho-sites
pSerine
pThreonine
pTyrosine
B
Mesenchymal
Epithelial
MDA-MB-231
MCF-7
1 2 3 4 5 6 7 8
Phospho-sites
TSC2
ENAH
MAP3K7
RALGPS2
NEDD4L
ROBO1
SPTAN1
MINK1
ARHGAP17
TBC1D23
AIB1
PLEKHA1
MPRIP
C
Motif
D
Phosphosites
Included in Epithelial-like cells
44
25
Skipped in epithelial-like cells
E
ESRP1/2_RBM47
MBNL1/2_RBFOX2
Phosphosites
Phospho-Serine
40
26
26
10
F
MCF-7
siESRP1&2
siRBM47
1 2 3
TSC2
ENAH
MAP3K7
RALGPS2
NEDD4L
ROBO1
SPTAN1
ARHGAP17
TBC1D23
KIAA0182
G
SC79
siESRP1&2
1 2 3 4
p4EBP1 (70)
p4EBP1 (36&47)
4EBP1
pS6
S6
ESRP1
ESRP2
H3
20
30
80
15
H
Normalized fold effect
1
0.5
p4EBP1 (70)
p4EBP1 (36&47)
pS6
SC79
siESRP1&2
I
SC79
TOSS/siTSC2
1 2 3
p4EBP1 (70)
p4EBP1 (36&47)
H3
20
15


A
EO Score
8
4
CE
ASE
Mes-Epi exons
Structure
Secondary structure
IUPR
B
Phosphosites (28)
5
23
14
« IUPR » &
« Protein Binding » (37)
C
Mesenchymal
Epithelial
MDA-MB-231
MCF-7
1 2 3 4 5 6 7 8
Interaction with autophagic factors
KIAA00226
WDFY3
SPAG9
PLOD2
SNX14
KIAA0182
EXOC7
PPP3CB
FLNB
FNIP1
SPTAN1
EXOC1
D
MDA
siMBNL1&2 - + -
siRBFOX2 - - +
1 2 3
KIAA0226
WDFY3
SPAG9
PLOD2
SNX14
EXOC7
PPP3CB
EXOC1
E
KIAA0226
BECN1
EXOC1
WDFY3
EXOC4
ATG5
EXOC7
PLOD2
WDFY3
SPAG9
MAP1LC3A
MAP1LC3B
MAP1LC3C
SQSTM1
WDFY3
PPP3CB
GABARAP
GABARAPL1
GABARAPL2
F
EBSS - + +
siM+R - - +
1 2 3
p62
LC3
LC3-I
LC3-II
MBNL1
MBNL2
RBFOX2
H3
60
40
40
50
15
Normalized p62 protein level
1
0.5
*
Normalized LC3 mRNA level
1.5
1
0.5
*
EBSS + +
siM+R - +
G
EBSS - + +
TOSS/siKIAA0226 - - +
1 2 3
p62
LC3
LC3-I
LC3-II
H3
60
15
H
TOSS/siKIAA0226 - +
MBNL1
MBNL2
H3
40
40
15
MBNL1 MBNL2
Normalized MBNL1&2 mRNA level
1.5
1
0.5
CTRL
TOSS/siKIAA0226
I
CTRL
TOSS/siKIAA0226
CTRL
siMBNL1&2
1 2 3 4
WDFY3
68 59 78 71
SPAG9
57 35 78 40
PLOD2
81 65 77 38
NFYA
71 52 77 35
CLSTN1
36 17 42 2
EXOC1
21 36 39 75

Bibliographie

[1000 Genomes Project Consortium et al., 2012] 1000 Genomes Project Consortium, T. . G. P., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422) :56–65.

[Ameur et al., 2010] Ameur, A., Wetterbom, A., Feuk, L., and Gyllensten, U. (2010). Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.*, 11(3) :R34.

[Anamika et al., 2016] Anamika, K., Verma, S., Jere, A., and Desai, A. (2016). Transcriptomic Profiling Using Next Generation Sequencing -Advances, Advantages, and Challenges.

[Anders et al., 2015] Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2) :166–169.

[Anders et al., 2012] Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10) :2008–17.

[Au et al., 2010] Au, K. F., Jiang, H., Lin, L., Xing, Y., and Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, 38(14) :4570–8.

[Bao et al., 2009] Bao, H., Xiong, Y., Guo, H., Zhou, R., Lu, X., Yang, Z., Zhong, Y., and Shi, S. (2009). MapNext : a software tool for spliced and unspliced alignments and SNP detection of short sequence reads. *BMC Genomics*, 10 Suppl 3(Suppl 3) :S13.

[Barash et al., 2010] Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. (2010). Deciphering the splicing code. *Nature*, 465(7294) :53–59.

[Bates et al., 2002] Bates, D. O., Cui, T.-G., Doughty, J. M., Winkler, M., Sugiono, M., Shields, J. D., Peat, D., Gillatt, D., and Harper, S. J. (2002). VEGF165b, an inhibitory splice variant of vascular endothelial growth factor, is down-regulated in renal cell carcinoma. *Cancer Res.*, 62(14) :4123–31.

[Bentley, 2002] Bentley, D. (2002). The mRNA assembly line : transcription and processing machines in the same factory. *Curr. Opin. Cell Biol.*, 14(3) :336–342.

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

[Bernard et al., 2014] Bernard, E., Jacob, L., Mairal, J., and Vert, J.-P. (2014). Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, 30(17) :2447–55.

[Bolger et al., 2014] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15) :2114–20.

[Bryant et al., 2010] Bryant, D. W., Shen, R., Priest, H. D., Wong, W.-K., and Mockler, T. C. (2010). Supersplat–spliced RNA-seq alignment. *Bioinformatics*, 26(12) :1500–5.

[Buenrostro et al., 2013] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12) :1213–8.

[Chi et al., 2009] Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254) :479–86.

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

[Chu et al., 2011] Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., and Chang, H. Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell*, 44(4) :667–78.

[Churchman and Weissman, 2011] Churchman, L. S. and Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330) :368–73.

[Cloonan et al., 2008] Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., and Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, 5(7) :613–619.

[Coopman et al., 2000] Coopman, P. J. P., Do, M. T. H., Barth, M., Bowden, E. T., Hayes, A. J., Basyuk, E., Blancato, J. K., Vezza, P. R., McLeskey, S. W., Mangeat, P. H., and Mueller, S. C.

(2000). The Syk tyrosine kinase suppresses malignant growth of human breast cancer cells. *Nature*, 406(6797) :742–747.

[Core et al., 2008] Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909) :1845–8.

[Cusack et al., 2011] Cusack, B. P., Arndt, P. F., Duret, L., and Roest Crollius, H. (2011). Preventing dangerous nonsense : selection for robustness to transcriptional error in human genes. *PLoS Genet.*, 7(10) :e1002276.

[Daemen et al., 2013] Daemen, A., Griffith, O. L., Heiser, L. M., Wang, N. J., Enache, O. M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J. E., Griffith, M., Hur, J. S., Huh, N., Chung, J., Cope, L., Fackler, M. J., Umbricht, C., Sukumar, S., Seth, P., Sukhatme, V. P., Jakkula, L. R., Lu, Y., Mills, G. B., Cho, R. J., Collisson, E. A., van't Veer, L. J., Spellman, P. T., and Gray, J. W. (2013). Modeling precision treatment of breast cancer. *Genome Biol.*, 14(10) :R110.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR : ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1) :15–21.

[Du et al., 2007] Du, L., Pollard, J. M., and Gatti, R. A. (2007). Correction of prototypic ATM splicing mutations and aberrant ATM function with antisense morpholino oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.*, 104(14) :6007–12.

[Dutertre et al., 2010] Dutertre, M., Sanchez, G., De Cian, M.-C., Barbier, J., Dardenne, E., Gratadou, L., Dujardin, G., Le Jossic-Corcos, C., Corcos, L., and Auboeuf, D. (2010). Cotranscriptional exon skipping in the genotoxic stress response. *Nat. Struct. Mol. Biol.*, 17(11) :1358–1366.

[Edery et al., 2011] 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., and Leutenegger, A.-L. (2011). Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. *Science*, 332(6026) :240–3.

[Feng et al., 2012] Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Shirley Liu, X., and Zhang, Y. (2012). GFOLD : a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, 28(21) :2782–2788.

[Friedman et al., 1999] Friedman, K. J., Kole, J., Cohn, J. A., Knowles, M. R., Silverman, L. M., and Kole, R. (1999). Correction of Aberrant Splicing of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Gene by Antisense Oligonucleotides. *J. Biol. Chem.*, 274(51) :36193–36199.

[Fu et al., 2013] Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., NHLBI Exome Sequencing Project, and Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431) :216–20.

[Fullwood and Ruan, 2009] Fullwood, M. J. and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, 107(1) :30–9.

[German et al., 2009] German, M. A., Luo, S., Schroth, G., Meyers, B. C., and Green, P. J. (2009). Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat. Protoc.*, 4(3) :356–362.

[Giresi and Lieb, 2009] Giresi, P. G. and Lieb, J. D. (2009). Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods*, 48(3) :233–9.

[Grabherr et al., 2011] 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., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7) :644–52.

[Grant et al., 2011] Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B., and Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18) :2518–28.

[Guttman et al., 2010] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28(5) :503–10.

[Guyot and Pagès, 2015] Guyot, M. and Pagès, G. (2015). VEGF Splicing and the Role of VEGF Splice Variants : From Physiological-Pathological Conditions to Specific Pre-mRNA Splicing.

In Fiedler, L., editor, *VEGF Signal. Methods Protoc.*, pages 3–23. Springer New York.

[Hafner et al., 2010] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1) :129–141.

[He et al., 2011] 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., and de la Chapelle, A. (2011). Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science*, 332(6026) :238–40.

[Hegele et al., 2012] Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C., Pena, V., Lührmann, R., and Stelzl, U. (2012). Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome. *Mol. Cell*, 45(4) :567–580.

[Huang et al., 2011] Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T.-W., Peng, Z., and Yiu, S.-M. (2011). SOAPsplice : Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Front. Genet.*, 2 :46.

[Huppertz et al., 2014] Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., and König, J. (2014). iCLIP : Protein–RNA interactions at nucleotide resolution. *Methods*, 65(3) :274–287.

[Iannone et al., 2015] Iannone, C., Pohl, A., Papasaikas, P., Soronellas, D., Vicent, G. P., Beato, M., and ValcáRcel, J. (2015). Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells. *RNA*, 21(3) :360–74.

[Ingolia et al., 2009] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924) :218–23.

[Ip et al., 2011] Ip, J. Y., Schmidt, D., Pan, Q., Ramani, A. K., Fraser, A. G., Odom, D. T., and Blencowe, B. J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res.*, 21(3) :390–401.

[Izquierdo et al., 2005] Izquierdo, J. M., Majós, N., Bonnal, S., Martínez, C., Castelo, R., Guigó, R., Bilbao, D., and Valcárcel, J. (2005). Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell*, 19(4) :475–84.

[Jacinto et al., 2008] Jacinto, F. V., Ballestar, E., and Esteller, M. (2008). Methyl-DNA immunoprecipitation (MeDIP) : hunting down the DNA methylome. *Biotechniques*, 44(1) :35, 37, 39 passim.

[Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830) :1497–502.

[Kalbfuss et al., 2001] Kalbfuss, B., Mabon, S. A., and Misteli, T. (2001). Correction of alternative splicing of tau in frontotemporal dementia and parkinsonism linked to chromosome 17. *J. Biol. Chem.*, 276(46) :42986–93.

[Katz et al., 2010] Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12) :1009–1015.

[Kent, 2002] Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4) :656–64.

[Kim et al., 2015] Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT : a fast spliced aligner with low memory requirements. *Nat. Methods*, 12(4) :357–60.

[Klinck et al., 2008] Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Gervais-Bird, J., Madden, R., Paquet, E. R., Koh, C., Venables, J. P., Prinos, P., Jilaveanu-Pelmus, M., Wellinger, R., Rancourt, C., Chabot, B., and Abou Elela, S. (2008). Multiple alternative splicing markers for ovarian cancer. *Cancer Res.*, 68(3) :657–63.

[Kornblihtt et al., 2013] Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing : a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.*, 14(3) :153–165.

[Kwak et al., 2013] Kwak, H., Fuda, N. J., Core, L. J., and Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122) :950–3.

[Lapuk et al., 2014] Lapuk, A. V., Volik, S. V., Wang, Y., and Collins, C. C. (2014). The role of mRNA splicing in prostate cancer. *Asian J. Androl.*, 16(4) :515–21.

[Lareau et al., 2007] Lareau, L. F., Brooks, A. N., Soergel, D. A. W., Meng, Q., and Brenner, S. E. (2007). The coupling of alternative splicing and nonsense-mediated mRNA decay. *Altern. SPLICING POSTGENOMIC ERA*, 623.

[Laustriat et al., 2015] Laustriat, D., Gide, J., Barrault, L., Chautard, E., Benoit, C., Auboeuf, D., Boland, A., Battail, C., ois Artiguenave, ois Deleuze, Rustin, P., Franc, S., Charpentier, G.,

Furling, D., Bassez, G., Nissan, X., cile Martinat, Peschanski, M., and Baghdoyan, S. (2015). In Vitro and In Vivo Modulation of Alternative Splicing by the Biguanide Metformin. *Cit. Mol. Ther. Acids*, 4.

[Lee and Wang, 2005] Lee, C. and Wang, Q. (2005). Bioinformatics analysis of alternative splicing. *Brief. Bioinform.*, 6(1) :23–33.

[Lee and Cooper, 2009] Lee, J. E. and Cooper, T. A. (2009). Pathogenic mechanisms of myotonic dystrophy. *Biochem. Soc. Trans.*, 37(Pt 6) :1281–6.

[Lewis et al., 2003] Lewis, B. P., Green, R. E., and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U. S. A.*, 100(1) :189–92.

[Lieberman-Aiden et al., 2009] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950) :289–93.

[Liu et al., 2013] Liu, Y., Ferguson, J. F., Xue, C., Silverman, I. M., Gregory, B., Reilly, M. P., and Li, M. (2013). Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLoS One*, 8(6) :e66883.

[Lopez-Maestre et al., 2016] Lopez-Maestre, H., Brinza, L., Marchet, C., Kielbassa, J., Bastien, S., Boutigny, M., Monnin, D., Filali, A. E., Carareto, C. M., Vieira, C., Picard, F., Kremer, N., Vavre, F., Sagot, M.-F., and Lacroix, V. (2016). SNP calling from RNA-seq data without a reference genome : identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res.*, page gkw655.

[Lou et al., 2011] Lou, S.-K., Ni, B., Lo, L.-Y., Tsui, S. K.-W., Chan, T.-F., and Leung, K.-S. (2011). ABMapper : a suffix array-based tool for multi-location searching and splice-junction mapping. *Bioinformatics*, 27(3) :421–2.

[Love et al., 2014] Love, M. I., Huber, W., Anders, S., Lönnstedt, I., Speed, T., Robinson, M., Smyth, G., McCarthy, D., Chen, Y., Smyth, G., Anders, S., Huber, W., Zhou, Y.-H., Xia, K., Wright, F., Wu, H., Wang, C., Wu, Z., Hardcastle, T., Kelly, K., Wiel, M. V. D., Leday, G., Pardo, L., Rue, H., Vaart, A. V. D., Wieringen, W. V., Boer, J., Huber, W., Sültmann, H., Wilmer, F., von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Füzesi, L.,

Vingron, M., Poustka, A., Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., Zhang, J., McCullagh, P., Nelder, J., Hansen, K., Irizarry, R., Wu, Z., Risso, D., Schwartz, K., Sherlock, G., Dudoit, S., Smyth, G., Bottomly, D., Walter, N., Hunter, J., Darakjian, P., Kawane, S., Buck, K., Searles, R., Mooney, M., McWeeney, S., Hitzemann, R., Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., Pritchard, J., Hastie, T., Tibshirani, R., Friedman, J., Bi, Y., Davuluri, R., Feng, J., Meyer, C., Wang, Q., Liu, J., Liu, X., Zhang, Y., Benjamini, Y., Hochberg, Y., Bourgon, R., Gentleman, R., Huber, W., McCarthy, D., Smyth, G., Li, J., Tibshirani, R., Cook, R., Hammer, P., Banck, M., Amberg, R., Wang, C., Petznick, G., Luo, S., Khrebtukova, I., Schroth, G., Beyerlein, P., Beutler, A., Frazee, A., Langmead, B., Leek, J., Trapnell, C., Hendrickson, D., Sauvageau, M., Goff, L., Rinn, J., Pachter, L., Glaus, P., Honkela, A., Rattray, M., Anders, S., Reyes, A., Huber, W., Sammeth, M., Robinson, M., McCarthy, D., Smyth, G., Zhou, X., Lindsay, H., Robinson, M., Leng, N., Dawson, J., Thomson, J., Ruotti, V., Rissman, A., Smits, B., Haag, J., Gould, M., Stewart, R., Kendziorski, C., Law, C., Chen, Y., Shi, W., Smyth, G., Hubert, L., Arabie, P., Witten, D., Irizarry, R., Wu, Z., Jaffee, H., Asangani, I., Dommeti, V., Wang, X., Malik, R., Cieslik, M., Yang, R., Escara-Wilke, J., Wilder-Romans, K., Dhanireddy, S., Engelke, C., Iyer, M., Jing, X., Wu, Y.-M., Cao, X., Qin, Z., Wang, S., Feng, F., Chinnaiyan, A., Ross-Innes, C., Stark, R., Teschendorff, A., Holmes, K., Ali, H., Dunning, M., Brown, G., Gojis, O., Ellis, I., Green, A., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C., Carroll, J., Robinson, D., Chen, W., Storey, J., Gresham, D., McMurdie, P., Holmes, S., Vasquez, J., Hon, C., Vanselow, J., Schlosser, A., Siegel, T., Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., Wei, W., Cox, D., Reid, N., Robinson, M., Smyth, G., Pawitan, Y., Armijo, L., Di, Y., Schafer, D., Cumbie, J., Chang, J., Abramowitz, M., Stegun, I., Newton, M., Kendziorski, C., Richmond, C., Blattner, F., Tsui, K., Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., Vingron, M., Durbin, B., Hardin, J., Hawkins, D., Rocke, D., Friedman, J., Hastie, T., Tibshirani, R., Cule, E., Vineis, P., Iorio, M. D., Cook, R., Weisberg, S., Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., Carey, V., Anders, S., Pyl, P., Huber, W., Delhomme, N., Padioleau, I., Furlong, E., Steinmetz, L., Liao, Y., Smyth, G., Shi, W., Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. (2014). Moderated estimation of fold

change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12) :550.

[Luco et al., 2010] Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science*, 327(5968) :996–1000.

[Mallinjoud et al., 2014] Mallinjoud, P., Villemin, J.-P., Mortada, H., Polay Espinoza, M., Desmet, F.-O., Samaan, S., Chautard, E., Tranchevent, L.-C., and Auboeuf, D. (2014). Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome Res.*, 24(3) :511–21.

[Marco-Sola et al., 2012] Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper : fast, accurate and versatile alignment by filtration. *Nat. Methods*, 9(12) :1185–1188.

[Martin et al., 2010] Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator : an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11 :663.

[Martin, 2011] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1) :pp. 10–12.

[Matlin et al., 2005] Matlin, A. J., Clark, F., and Smith, C. W. J. (2005). Understanding alternative splicing : towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6(5) :386–398.

[Maunakea et al., 2010] Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J. M., Haussler, D., Marra, M. A., Hirst, M., Wang, T., and Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303) :253–7.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nat. Rev. Genet.*, 11(1) :31–46.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7) :621–628.

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

[Orian-Rousseau, 2015] Orian-Rousseau, V. (2015). CD44 Acts as a Signaling Platform Controlling Tumor Progression and Metastasis. *Front. Immunol.*, 6 :154.

[Pan et al., 2008] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12) :1413–1415.

[Pérez et al., 2010] Pérez, B., Rodríguez-Pascau, L., Vilageliu, L., Grinberg, D., Ugarte, M., and Desviat, L. R. (2010). Present and future of antisense therapy for splicing modulation in inherited metabolic disease. *J. Inherit. Metab. Dis.*, 33(4) :397–403.

[Pertea et al., 2015] Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33(3) :290–5.

[Pico et al., 2008] Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). WikiPathways : Pathway Editing for the People. *PLoS Biol.*, 6(7) :e184.

[Prochazka et al., 2014] Prochazka, L., Tesarik, R., and Turanek, J. (2014). Regulation of alternative splicing of CD44 in cancer. *Cell. Signal.*, 26(10) :2234–2239.

[Pros et al., 2009] Pros, E., Fernández-Rodríguez, J., Canet, B., Benito, L., Sánchez, A., Benavides, A., Ramos, F. J., López-Ariztegui, M. A., Capellá, G., Blanco, I., Serra, E., and Lázaro, C. (2009). Antisense therapeutics for neurofibromatosis type 1 caused by deep intronic mutations. *Hum. Mutat.*, 30(3) :454–462.

[Proudfoot and O'Sullivan, 2002] Proudfoot, N. and O'Sullivan, J. (2002). Polyadenylation : A tail of two complexes. *Curr. Biol.*, 12(24) :R855–R857.

[Reuter et al., 2015] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell*, 58(4) :586–97.

[Reynoso et al., 2015] Reynoso, M. A., Juntawong, P., Lancia, M., Blanco, F. A., Bailey-Serres, J., and Zanetti, M. E. (2015). Translating Ribosome Affinity Purification (TRAP) Followed by RNA Sequencing Technology (TRAP-SEQ) for Quantitative Assessment of Plant Translatomes. In *Plant Funct. Genomics - Methods Protoc.*, chapter Part II, pages 185–207. Springer New York.

[Robertson et al., 2010] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B.,

Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A., and Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nat. Methods*, 7(11) :909–912.

[Rothberg et al., 2011] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T., and Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356) :348–352.

[Ruby et al., 2006] Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6) :1193–207.

[Ryan et al., 2012] Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., and Weinstein, J. N. (2012). SpliceSeq : a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, 28(18) :2385–7.

[Sacomoto et al., 2012] Sacomoto, G. A. T., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., and Lacroix, V. (2012). KISSPLICE : de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, 13 Suppl 6 :S5.

[Sahin et al., 2009] Sahin, O., Fröhlich, H., Löbke, C., Korf, U., Burmester, S., Majety, M., Mattern, J., Schupp, I., Chaouiya, C., Thieffry, D., Poustka, A., Wiemann, S., Beissbarth, T., and Arlt, D. (2009). Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC Syst. Biol.*, 3 :1.

[Sameen et al., 2016] Sameen, S., Barbuti, R., Milazzo, P., Cerone, A., Del Re, M., and Danesi, R. (2016). Mathematical modeling of drug resistance due to KRAS mutation in colorectal cancer. *J. Theor. Biol.*, 389 :263–273.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12) :5463–7.

[Schmieder and Edwards, 2011] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6) :863–4.

[Schulz et al., 2012] Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases : robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8) :1086–92.

[Schwartz et al., 2009] Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.*, 16(9) :990–995.

[Serre et al., 2010] Serre, D., Lee, B. H., and Ting, A. H. (2010). MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.*, 38(2) :391–9.

[Shen et al., 2012] Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.-x., Zhou, Q., Carstens, R. P., and Xing, Y. (2012). MATS : a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.*, 40(8) :e61.

[Shen et al., 2014] Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS : robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.*, 111(51) :E5593–601.

[Shendure and Ji, 2008] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10) :1135–1145.

[Singh and Cooper, 2012] Singh, R. K. and Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.*, 18(8) :472–82.

[Song and Crawford, 2010] Song, L. and Crawford, G. E. (2010). DNase-seq : a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, 2010(2) :pdb.prot5384.

[Southwell et al., 2012] Southwell, A. L., Skotte, N. H., Bennett, C. F., Hayden, M. R., Bennett, C., Swayze, E., Cerritelli, S., Crouch, R., Mills, J., Janitz, M., Stein, C., et Al., Summerton, J., Nielsen, P., et Al., Teplova, M., et Al., Seth, P., et Al., Kumar, R., et Al., Senn, J., et Al., Krieg, A., ENCODE, Freier, S., Watt, A., Xu, Z., et Al., Chiang, M., et Al., Carroll, J., et Al., Basilion, J., et Al., Lima, W., et Al., Monia, B., et Al., Banks, W., et Al., Erickson, M., et Al., Swayze, E., et Al., Passini, M., et Al., Smith, R., et Al., Kordasiewicz, H., et Al., Benoist, M., et Al., Freise, H., Aken, H. V., Hayek, S., et Al., Ross, C., Tabrizi, S., Group, T. H. C. R., Yamamoto, A., et Al., Harper, S., et Al., Rodriguez-Lebron, E., et Al., Zuccato, C., et Al., Boado, R., et Al., Nellemann, C., et Al., Slow, E., et Al., Gray, M., et Al., Nasir, J., et Al., Dragatsis, I., et Al., Grondin, R., et Al., McBride, J., et Al., Boudreau, R., et Al., Hodgson,

J., et Al., Raamsdonk, J. V., et Al., Raamsdonk, J. V., et Al., Liu, W., et Al., Krzyzosiak, W., et Al., den Driessche, T. V., et Al., Hu, J., et Al., Gagnon, K., et Al., Fiszer, A., et Al., Evers, M., et Al., Hu, J., et Al., Butland, S., et Al., Warby, S., et Al., Pfister, E., et Al., Warby, S., et Al., Crowther, R., Goedert, M., Lee, G., et Al., Donahue, C., et Al., Kalbfuss, B., et Al., Finder, V., Oddo, S., et Al., James, F., Chauhan, N., Siegel, G., Iannaccone, S., Zhou, J., et Al., Lorson, C., et Al., Prior, T., et Al., Porensky, P., et Al., Williams, J., et Al., Hua, Y., et Al., Hua, Y., et Al., Singh, N., et Al., Hua, Y., et Al., Hua, Y., et Al., Hardiman, O., et Al., Kuzma-Kozakiewicz, M., Kwiecinski, H., Gurney, M., et Al., Reaume, A., et Al., Miller, T., et Al., Osman, E., et Al., Moulton, H., and Moulton, J. (2012). Antisense oligonucleotide therapeutics for inherited neurodegenerative diseases. *Trends Mol. Med.*, 18(11) :634–43.

[Spies et al., 2009] Spies, N., Nielsen, C. B., Padgett, R. A., and Burge, C. B. (2009). Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, 36(2) :245–54.

[Srebrow and Kornblihtt, 2006] Srebrow, A. and Kornblihtt, A. R. (2006). The connection between splicing and cancer. *J. Cell Sci.*, 119(Pt 13) :2635–41.

[Sudmant et al., 2015] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571) :75–81.

[Tang and Riva, 2013] Tang, S. and Riva, A. (2013). PASTA : splice junction identification from RNA-sequencing data. *BMC Bioinformatics*, 14 :116.

[Tazi et al., 2009] Tazi, J., Bakkour, N., and Stamm, S. (2009). Alternative splicing and disease. *Biochim. Biophys. Acta - Mol. Basis Dis.*, 1792(1) :14–26.

[Tennessen et al., 2012] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad GO, Seattle GO, and NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090) :64–9.

[Terns and Terns, 2001] Terns, M. P. and Terns, R. M. (2001). Macromolecular complexes : SMN — the master assembler. *Curr. Biol.*, 11(21) :R862–R864.

[Tilgner et al., 2009] Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.*, 16(9) :996–1001.

[Toyama et al., 2003] Toyama, T., Iwase, H., Yamashita, H., Hara, Y., Omoto, Y., Sugiura, H., Zhang, Z., and Fujii, Y. (2003). Reduced expression of the Syk gene is correlated with poor prognosis in human breast cancer. *Cancer Lett.*, 189(1) :97–102.

[Trapnell et al., 2013] Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, 31(1) :46–53.

[Trapnell et al., 2009] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat : discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9) :1105–11.

[Trapnell et al., 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5) :511–515.

[Tsompana and Buck, 2014] Tsompana, M. and Buck, M. J. (2014). Chromatin accessibility : a window into the genome. *Epigenetics Chromatin*, 7(1) :33.

[Vargas et al., 2011] Vargas, D. Y., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S. A. E., Schedl, P., and Tyagi, S. (2011). Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, 147(5) :1054–65.

[Venables et al., 2008] Venables, J. P., Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Koh, C., Gervais-Bird, J., Lapointe, E., Froehlich, U., Durand, M., Gendron, D., Brosseau, J.-P., Thibault, P., Lucier, J.-F., Tremblay, K., Prinos, P., Wellinger, R. J., Chabot, B., Rancourt,

C., and Elela, S. A. (2008). Identification of alternative splicing markers for breast cancer. *Cancer Res.*, 68(22) :9525–31.

[Wahl et al., 2009] Wahl, M. C., Will, C. L., and Lührmann, R. (2009). The spliceosome : design principles of a dynamic RNP machine. *Cell*, 136(4) :701–18.

[Wang and Cooper, 2007] Wang, G.-S. and Cooper, T. A. (2007). Splicing in disease : disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, 8(10) :749–761.

[Wang et al., 2010] Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., and Liu, J. (2010). MapSplice : accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38(18) :e178.

[Wang et al., 2003] Wang, L., Duke, L., Zhang, P. S., Arlinghaus, R. B., Symmans, W. F., Sahin, A., Mendez, R., and Dai, J. L. (2003). Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer Res.*, 63(15) :4724–30.

[Wang et al., 2011] Wang, L., Wang, X., Wang, X., Liang, Y., and Zhang, X. (2011). Observations on novel splice junctions from RNA sequencing data. *Biochem. Biophys. Res. Commun.*, 409(2) :299–303.

[Weischenfeldt et al., 2012] Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J. S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B. T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biol.*, 13(5) :R35.

[Wu and Nacu, 2010] Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7) :873–81.

[Wu and Watanabe, 2005] Wu, T. D. and Watanabe, C. K. (2005). GMAP : a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9) :1859–1875.

[Xerri et al., 1998] Xerri, L., Bouabdallah, R., Devilard, E., Hassoun, J., Stoppa, A. M., and Birg, F. (1998). Sensitivity to Fas-mediated apoptosis is null or weak in B-cell non-Hodgkin's lymphomas and is moderately increased by CD40 ligation. *Br. J. Cancer*, 78(2) :225–32.

[Yeo and Burge, 2004] Yeo, G. and Burge, C. B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.*, 11(2-3) :377–394.

[Yoshida and Ogawa, 2014] Yoshida, K. and Ogawa, S. (2014). Splicing factor mutations and cancer. *Wiley Interdiscip. Rev. RNA*, 5(4) :445–459.

[Zhang et al., 2012] Zhang, Y., Lameijer, E.-W., 't Hoen, P. A. C., Ning, Z., Slagboom, P. E., and Ye, K. (2012). PASSion : a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics*, 28(4) :479–86.