Université de Lille

Thèse

pour obtenir le grade de :

Docteur de l'Université de Lille

dans la spécialité

« Informatique »

par

Lilian Marchand

Méthodes pour la reconstruction du répertoire des transcrits d'un gène à partir de données RNA-seq de 3ème génération

Thèse soutenue le 18 octobre 2024 devant le jury composé de :

Mme	Aïda OUANGRAOUA	Professeure des universités, Université de Sherbrooke	(Rapportrice)
Mme	Elodie LAINE	Professeure des universités, Sorbone Université, LCQB	(Rapportrice)
M.	François BOULIER	Professeur des universités, Université de Lille, CRIStAL	(Président du jury)
M.	Vincent LACROIX	Maître de conférences, Université Lyon 1, LBBE	(Examinateur)
Mme	Hélène TOUZET	Directrice de recherche, CNRS, CRIStAL	(Invitée)
M.	Jean-Stéphane VARRE	Professeur des universités, Université de Lille, CRIStAL	(Directeur de Thèse)

Laboratoire CRIStAL UMR 9189

Remerciements

Je tiens tout d'abord à remercier sincèrement les membres du jury : Élodie Laine, Aïda Ouangraoua, Vincent Lacroix, et François Boulier, pour avoir accepté de participer à ma soutenance. Un merci tout particulier à Aïda et Élodie pour le temps qu'elles ont consacré à la relecture et à l'évaluation de mon manuscrit, ainsi qu'à Élodie pour la relecture de mon CSI.

Je souhaite exprimer toute ma gratitude à mon directeur, Jean-Stéphane, pour sa bienveillance, son soutien constant et son aide précieuse au cours de ces trois années. Merci aussi à Hélène pour le temps qu'elle a passé à réfléchir avec nous sur ce sujet.

Je tiens également à remercier tous les membres de l'équipe Bonsaï que j'ai eu la chance de côtoyer pendant ces trois années de thèse et avec qui j'ai partagé de nombreux bureaux. Merci à Coralie de nous avoir montré la voie, à Thomas pour sa passion communicative des jeux de stratégie, à Pierre G. pour nos discussions enrichissantes, et à Léa pour avoir incarnée la culture du Nord. Un grand merci aussi à tous les nouveaux arrivants : Timothée, Pierre, Igor, Caleb, Madeleine, Florian, pour le vent de fraîcheur et la bienveillance qu'ils ont apportés.

Merci également à Bastien, Areski, Michael, Camille et Antoine pour leurs précieux conseils et leur hospitalité. Merci à Laurent pour sa présentation claire des graines espacées.

Un merci spécial à Caleb pour ses récaps des étapes du Tour de France et de m'avoir transmis la délicieuse recette des scones de sa maman.

Merci également à tous les participants du futsal de l'université, les bons comme les mauvais perdants, pour ces rendez-vous hebdomadaires au Cosec qui étaient une vraie bouffée d'air frais.

Un immense merci à Cécile, qui partage ma vie, pour sa gentillesse, son soutien infaillible au quotidien, ses encouragements et son amour.

Je souhaite aussi remercier Ludovic, Aurore et Malika à la direction de l'ED MADIS pour leur accompagnement dans cette expérience très enrichissante. Un grand merci également à Cédric pour sa complicité.

Enfin, merci aux membres du conseil d'unité, qui m'ont permis de découvrir le fonctionnement interne de celui-ci.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

L'épissage alternatif est un processus de régulation de la structure interne des ARNs, contribuant à la maturation des Pré-ARN en ARN matures. Ce processus, synchrone à la transcription, permet à la cellule de produire différents ARN à partir d'un même gène et donc d'augmenter la diversité protéique. On estime que 95% des gènes humains connaissent des événements d'épissage alternatif, impliqués dans des processus cellulaires cruciaux tels que la différenciation et le développement cellulaire, la réponse au stress environnemental et l'immunité [RQQ19]. Sa grande diversité rend presque impossible la quête d'exhaustivité dans l'identification de ces formes, tant ces variations peuvent être circonscrites dans le temps et dans l'espace[BG17]. Ainsi son étude est intéressante pour répondre à de nombreuses questions biologiques. Son importance va de pair avec les conséquences pathologiques de l'arrivée d'un problème dans le processus d'épissage. Il a été identifié comme la cause de nombreuses maladies graves, dont les cancers [Bes+20]. Être en capacité de répertorier précisément les transcrits alternatifs dans ces cas est important pour identifier des cibles pour réaliser des traitements précis.

L'épissage alternatif est aujourd'hui étudié grâce au séquençage du transcriptome, dit de seconde ou de troisième génération. Les technologies de seconde génération, avec des lectures courtes (150 à 300 pb), offrent une excellente qualité de séquençage (taux d'erreur inférieur à 0,01 %). Cependant, la longueur limitée des lectures empêche le séquençage des transcrits en pleine longueur, entraînant une détermination ambiguë des combinaisons d'exons au sein des isoformes. À l'inverse, le séquençage de troisième génération propose des lectures longues couvrant la totalité de la longueur du transcrit (jusqu'à 30 kb), mais avec un taux d'erreur plus élevé (de 5 à 10 %) et une tendance à la troncation en début et en fin de lectures [Wan+21]. Travailler avec des lectures longues nécessite la prise en compte de différents problèmes comme les erreurs de séquençage, la dégradation en début et fin de lectures mais aussi l'intégrité des lectures et la faible profondeur de séquençage [Byr+19].

Depuis le début des années 2020, l'importance de l'épissage alternatif et le développement continu des technologies de séquençage de 3ème génération ont conduit à une augmentation des publications d'outils d'identification des isoformes alternatifs. Ces outils proposent diverses approches pour identifier les isoformes d'épissage avec [Tan+20 ; Gao+23 ; Prj+23] ou sans séquence de référence [PS23] et avec ou sans annotation [Ora+23]. Cependant, ils se placent toujours à l'échelle du génome, ce qui implique des contraintes fortes d'efficacité en temps étant donné la grande taille des jeux de données longues lectures.

En contrepied de cette démarche, la question de recherche abordée dans cette thèse est de savoir *comment être capable d'identifier à partir d'une expérience de transcriptomique de troisième génération et pour un gène donné l'ensemble des ARNm variants d'épissage que ce gène peut produire*. Cela sous-tend de développer les approches méthodologiques nécessaires pour être le plus précis possible et faciliter une recherche exploratoire au sein des isoformes putatifs prédits par un outil répondant à cette question. En réponse à cette

question nous présentons dans cette thèse le développement de RNA-tailor. Il est le premier outil d'identification d'isoformes alternatifs fonctionnant sans annotation et à l'échelle du gène. Fonctionner sans annotation est un avantage pour identifier l'ensemble des isoformes sans a priori. Fonctionner à l'échelle du gène présente l'avantage de pouvoir diminuer la pression d'optimisation sur le temps de calcul et ainsi de s'ouvrir à une démarche plus exploratoire. Cela en fait un outil versatile pour l'étude des événements d'épissage de novo, chez des espèces modèles ou non modèles, sans connaissance a priori autre qu'une séquence génomique de référence.

Nous présentons en chapitre 2 un état de l'art nécessaire pour comprendre les mécanismes moléculaires de l'épissage et sa régulation dans la cellule. On y décrit les technologies de séquençage existantes et les challenges que représentent leur utilisation dans le cadre de la transcriptomique. Puis on présente différents outils déjà publiés de prédictions d'isoformes à partir de données de séquençage de troisième génération.

Dans le chapitre 3, on exposera notre contribution méthodologique à la problématique de recherche, en expliquant en détails chaque étape de la méthode et les outils utilisés pour la conception de RNA-tailor. Nous décrirons son pipeline d'analyse, passant par plusieurs étapes de sélection, d'alignements, de filtrage et de correction, conçu pour identifier précisément les variants d'épissage à l'échelle d'un gène.

Dans le chapitre 4, on compare les performances de prédiction de RNA-tailor avec d'autres outils de l'état de l'art. Pour cela, on choisit d'évaluer la qualité de prédictions des outils sur des données simulées et réelles ainsi que la capacité de ces derniers à reconnaître des événements d'épissage artificiels. Dans le cadre de cette comparaison, se pose la problématique de savoir quel aspect des résultats doit être pris en compte : faut-il évaluer la prédiction sur l'ensemble du transcrit complet ou se concentrer uniquement sur sa structure intronique ? L'identification précise des sites d'épissage constitue une première approche pour aborder cette question. Ces résultats permettent d'apprécier la difficulté de la tâche de prédiction d'isoformes et pointent du doigt l'absence de consensus entre méthodes.

Enfin, dans le chapitre 5, nous exposons nos conclusions sur ce travail et les perspectives méthodologiques pour le développement de RNA-tailor. On y mentionne les moyens pour continuer à améliorer ses performance mais aussi les perspectives d'utilisation alternative de RNA-tailor, notamment la potentialité d'utiliser une séquence de référence qui ne soit pas celle de l'espèce sur laquelle le transcriptome a été séquencé.

Chapitre 2

Éléments Bibliographiques

Sommaire

Introduction

Au début des années 2000, le dogme central de la biologie moléculaire soutenait que chaque gène codait une unique protéine. La découverte de l'existence d'isoformes alternatifs a remis en question la règle « un gène, une protéine, une fonction ». L'avènement du séquençage génomique à grande échelle a permis d'avoir les outils nécessaires à son étude. L'évolution des technologies de séquençage, passant de lectures courtes à des lectures plus longues, a considérablement amélioré notre capacité à détecter ces variations. Aujourd'hui, il est évident que les isoformes alternatifs jouent un rôle crucial dans de nombreux mécanismes biologiques. Les technologies actuelles ne permettent pas d'identifier tous les isoformes possibles. Dans ce domaine, la course à l'exhaustivité est impossible tant la production d'isoformes peut être courte dans le temps et spécifique à certains tissus et/ou conditions biologiques. Cette partie présente un cadre bibliographique qui explore les bases de l'épissage alternatif, les progrès des technologies de séquençage, et les méthodes pour analyser et évaluer les variants d'épissage grâce au séquençage de troisième génération.

2.1 Mécanismes et fonctions de l'épissage alternatif

2.1.1 De la transcription à la production de transcrits matures

Structure interne des gènes

Un gène est dit exprimé si une molécule d'ARN est produite par la transcription de son locus. Il existe différents niveaux de régulation de l'expression des gènes. Certains sont intégrés dans la structure de l'ADN sur le même chromosome que le gène, c'est le cas des

séquences promotrices de la transcription ou promoteurs. Ces promoteurs comportent des séquences conservées comme la *CAAT box* et la *TATA box* qui se situent respectivement 70nt et 25nt en amont du début de la séquence codante du gène. Ce sont des points de fixation de l'ARN polymérase et de facteurs de transcriptions. Les facteurs de transcriptions peuvent avoir un effet activateur (*enhancer*) ou inhibiteur (*silencer*) sur l'activité et la bonne fixation de l'ARN polymérase. Ces facteurs de transcriptions protéiques peuvent être le produit de chaînes d'activations enclenchées par un signal hormonal ou transmembranaire. La structure interne des gènes eucaryotes est présentée Figure 2.1.

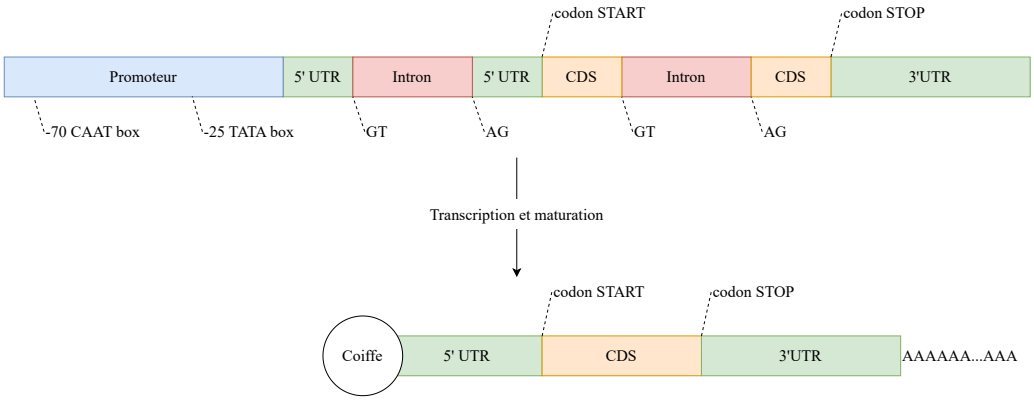


FIGURE 2.1 – Schéma de la structure interne d'un gène eucaryote et d'un ARN mature produit après transcription du gène et après maturation de l'ARNm.

La transcription de l'ADN en ARN par l'ARN polymérase se fait grâce à la complémentarité des bases. La molécule d'ARN est composé de quatre bases : l'adénine (A), la guanine (G), la cytosine (C) et l'uracile (U), unique à l'ARN. L'information traductionnelle portée par l'ARN est organisée en codons de trois bases et est traduite en acide aminé par le code génétique. La transcription commence avec la reconnaissance d'un codon d'initiation START (AUG) et finit avec un codon terminal STOP (UAA, UAG et UGA). On parle de aussi de TTS et TSS, respectivement *Transcription Termination Site* et *Transcription Start Site*.

L'épissage alternatif

L'épissage alternatif ou *alternative splicing* (AS) est un mécanisme de régulation de la structure des ARN chez les eucaryotes. Les gènes sont structurés par des parties de séquences codantes (exons) et non-codantes (introns). Afin que la transcription d'un gène par l'ARN-polymérase puisse mener à la production d'une protéine viable, seuls les exons doivent être conservés. Cependant, l'ensemble des exons initiaux ne sont pas nécessairement conservés dans un ARNm prêt à être transcrit. Cette étape d'épuration de la séquence brute du pré-ARN est appelée épissage, ou *splicing*. Pour qualifier les ARNm produits par l'épissage, on parle de variants d'épissage ou de transcripts isoformes. Par abus de langage, on utilise parfois le terme *isoforme* seul, il est dans ce cas bien destiné à parler des transcripts en non des protéines. L'épissage alternatif est un phénomène largement répandu chez l'humain. Certaines études estiment que 95% des gènes présentent des événements d'épissage alternatif [Bes+20]. Mais ce mécanisme n'est pas restreint à l'humain, il a également été mis en évidence chez les plantes et les poissons [BG17].

L'épissage est un vecteur de diversité de production des transcrits [KLA10]. Ce mécanisme permet de tirer parti de la combinatoire des différents sous-parties codantes d'un gène. En effet, un ARN mature (ARNm) n'a pas forcément besoin d'inclure tous les exons du gène pour être viable. Nous discutons un peu plus loin des mécanismes d'épissage et des différents évènements d'épissage alternatif.

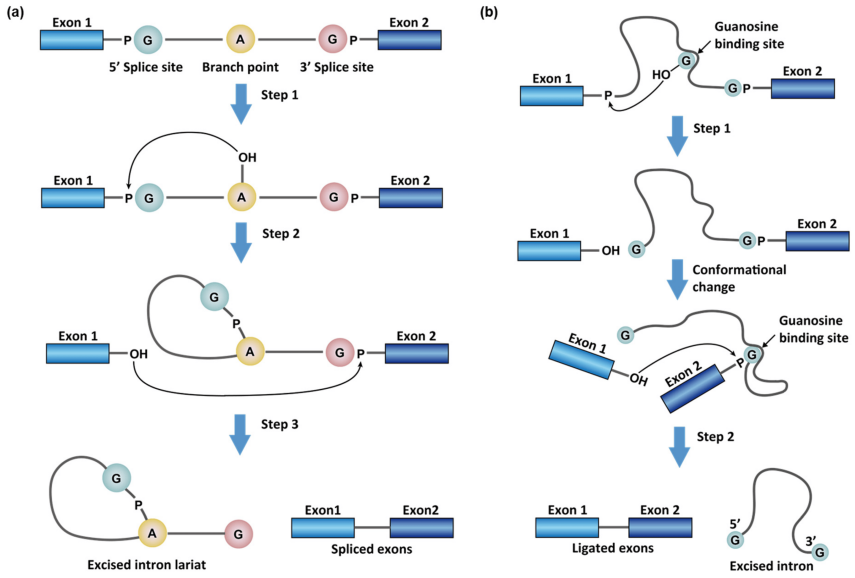


FIGURE 2.2 – Schéma des étapes des deux chaînes de réactions possibles menant à l'excision d'un intron. La chaîne de réaction (a) correspond à la chaîne de réactions catalysées par le spliceosome ou pour les introns de groupe II. La chaîne de réaction (b) correspond à la chaîne de réactions catalysées pour les introns de groupe I., tiré de [Tan+21].

A titre d'exemple, nous avons répertorié quelques données sur les annotations de gènes dans la version 111 du génome humain GRCh38 (Tableau 2.1).

Métrique	**Valeur**
Nombre de gènes	20073
Nombre de transcrits codant pour une protéine	170505
Nombre moyen de transcrits par gène	8.4
Nombre moyen d'exons par transcrit	8.1
Nombre total de combinaisons d'exons possibles	5169615
Pourcentage des transcrits observés en comparaison	3.2%

TABLE 2.1 – Résumé des données de transcrits du génome humain version 111 GRCh38. Le nombre de combinaisons d'exons est donné en supposant 8 exons par gène, toute combinaison possible entre 1 et 8 exons pour chaque gène, soit $20073 \times (2^8 - 1)$

On compte 170505 transcripts avec le tags *protein coding* pour 20073 gènes, soit une moyenne de plus de 8 transcripts par gène. Chaque transcrit possède en moyenne un peu plus de 8 exons.Cette diversité d'exons par gène serait suffisante pour générer plus de 5 millions de combinaisons d'exons différentes (en utilisant les moyennes). Cela ne représente que 3.2% des transcrits possiblement réalisables. La diversité réellement exprimée est donc très faible en regard des possibles. Cela provient notamment du fait que le spliceosome sélectionne des combinaisons d'exons spécifiques, et de l'existence de mécanismes comme le complexe NMD dégradant les transcrits identifiés comme contaminants.

Mécanismes d'épissage du spliceosome

Le splicing est effectué par le *spliceosome*. C'est un complexe ribonucléoprotéique (RNP) composé de nombreuses protéines (> 200) [Heg+12]. Il reconnaît les sites d'épissage et effectue l'épissage des introns (voir Figure 2.2a).

Il enlève les séquences introniques non-codantes des ARN messager-précurseurs [Heg+12]. Sa nature dynamique, tant compositionnelle que structurelle lui confère deux caractéristiques importantes du point de vue biologique : une flexibilité en termes de choix d'épissage

et une bonne précision de reconnaissance des loci d'épissage. Ainsi le répertoire particulier de protéines présentes dans le spliceosome à chaque étape du cycle de vie de la cellule, détermine le devenir d'un pré-ARNm [WWL09]. Bien que l'épissage par le spliceosome soit prédominant chez les eucaryotes, il existe également des types d'introns qui n'ont pas besoin du spliceosome pour être épissés. C'est le cas des introns dit *self-splicing* [Tan+21]. Ils sont caractérisés en tant qu'introns de groupe I et II que l'on trouve dans les parties codantes et non-codantes de l'ARN dans tout le règne du Vivant (voir Figure 2.2). Néanmoins ce type d'épissage n'est jamais observé chez les animaux, mais largement répandu chez les procaryotes. L'épissage est un événement synchrone à la transcription de l'ARN. Ainsi, dans les résultats de séquençage, il est possible d'identifier des transcrits à différentes étapes d'épissage. Dans certains cas, il est même possible d'observer la conservation d'un ordre d'épissage des introns [Cho+23].

L'épissage par le spliceosome est illustré avec plus de détails Figure 2.3. La séparation

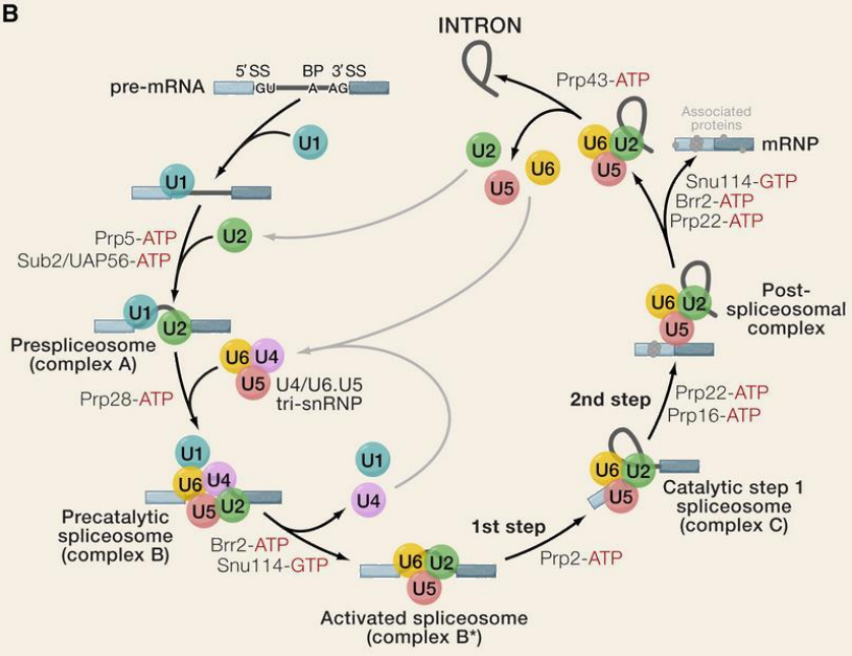


FIGURE 2.3 – Schéma du mécanisme d'épissage catalysé par le spliceosome. Tiré de [WWL09].

des parties introniques des parties exoniques est réalisée grâce à différentes sous-unités ribonucléoprotéiques (snRNP) du spliceosome. Dans le cadre de la thèse, on s'intéresse au spliceosome majeur, aussi appelé spliceosome U2. Les étapes de formation et de désolidarisation des snRNP de ce complexe correspondent aux grandes étapes du mécanisme d'épissage :

1. reconnaissance par les sous-unités U1 et U2 des sites d'épissage canoniques GT-AG et s'y fixer ;
2. liaison des sous-unité U4, U5 et U6 au précomplexe ARN-U1-U2 ;
3. activation du complexe et départ de U4 et U1 ;
4. catalyse de l'excision de l'intron ;
5. séparation du complexe U2-U6-U5 ;
6. libération du lasso d'intron.

Chez les animaux, les lassos d'introns produits par l'épissage sont dégradés en quelques minutes. Certains cependant restent à l'état cyclique et sont exportés dans le cytoplasme où ils peuvent réguler des fonctions cellulaires [Tan+21]. La sélection des bornes des introns est réalisée grâce à la reconnaissance d'un ensemble de motifs conservés au niveau des sites donneurs (GT ou GC) et accepteurs (AG) et dans la séquence de l'intron lui-même. Il y a notamment une adénine très conservée 20 nt en amont du site accepteur qui correspond au site de branchement de la sous unité U2 et de ces co-facteurs (Figure 2.4). Plus ces motifs

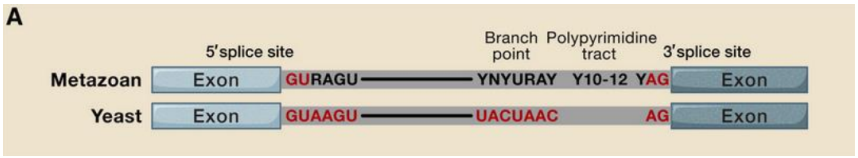


FIGURE 2.4 – Motifs conservés dans les introns. Tiré de [WWL09].

sont conservés et plus les sites respectifs d'épissage sont dit forts et vraisemblablement sélectionnés par le spliceosome.

La reconnaissance des sites d'épissage ne veut pas forcément dire que l'exon sera inclus dans l'isoforme final. Pour qu'un exon soit sélectionné, la balance de présence de différents facteurs d'épissage est mise en oeuvre. Parmi ces éléments, on trouve les *exonic splicing enhancer* (ESE) qui favorisent l'épissage, et les *exonic splicing silencer* (ESS) qui l'inhibent. De même, les *intronic splicing enhancer* (ISE) stimulent l'épissage, tandis que les *intronic splicing silencer* (ISS) le répriment. La figure 2.5 montre également l'influence de molécules d'ARN qui peuvent inhiber un site d'épissage en se liant en 5' et perturber la fixation de U2. Un site d'épissage est l'un des deux sites (donneur ou accepteur) d'une jonction d'épissage. Pour parler d'un couple de site d'épissage, on peut aussi parler de jonction intronique ou exonique.

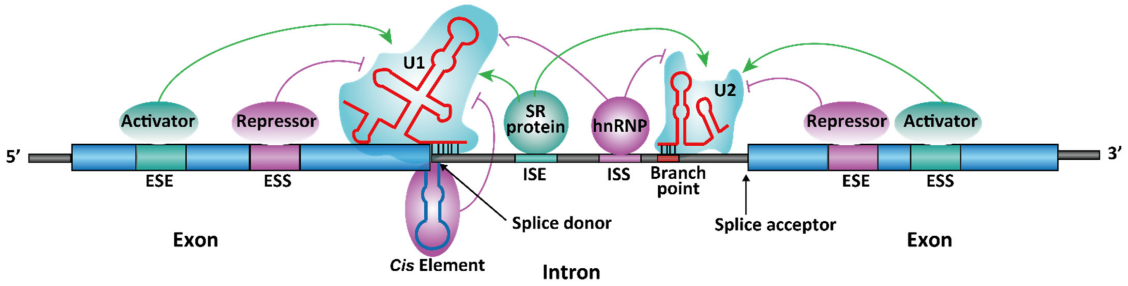


FIGURE 2.5 – Régulation de l'épissage par des co-facteurs. Tiré de [Tan+21].

Une fois les introns épissés, d'autres étapes de la maturation de l'ARN suivent. Il s'agit du phénomène de 5' capping et de polyadenylation en 3' de l'ARN qui vont lui donner les caractéristiques physico-chimiques d'un ARN messager mature. L'ARNm va pouvoir sortir du noyau pour être traduit en protéine par le ribosome.

Traduction des ARNm en protéine

Les ARN messagers codants peuvent ensuite être traduits en protéine. Celle-ci est composée d'une chaîne d'acides aminés. La séquence de cette chaîne est encodée par le code génétique et est assemblée par le ribosome (un complexe ribonucléoprotéique). Il reconnaît la séquence codante de l'ARNm grâce au codon d'initiation et de terminaison. Ainsi les parties non-codantes, UTR (UnTranslated Regions), même si elles sont présentes dans le transcrit, ne seront pas traduites.

Il existe également des ARNs qui ne sont pas traduits en protéines. Ces ARN non codants, appelés sncRNA (small non-coding RNA) et *lncRNA* (*long non-coding RNA*)

jouent néanmoins un rôle dans la régulation de l'expression des gènes et de processus cellulaires [MRH17].

Rôle du splicing

Rôle dans le développement Les réseaux de régulation de l'épissage alternatif permettent de coordonner le développement des organes et la mise en place des tissus différenciés chez les mammifères et les plantes [BG17]. Ainsi chez l'Homme, on peut prendre l'exemple de la Tropomyosine dont l'épissage va permettre la mise en place de différents type de muscles (squelettiques, lisses ou cardiaques). Chez les plantes, le gène ABI5 subit un épissage alternatif lors de la germination [SD23]. Cet épissage diminue la sensibilité de l'isoforme produit à l'acide abscissique, inhibiteur de la germination, et avec lui celle de la cellule. La graine devient moins sensible à l'hormone et la graine peut germer.

Les mécanismes d'épissage alternatif ont été identifiés dans de nombreux mécanismes biologiques importants. Ils sont impliqués dans l'horloge circadienne, la floraison, l'adaptation environnementale (stress abiotique) mais également aux stress pathogénique [BG17].

Rôle dans l'immunité. Chez l'Homme, dans le cas d'une infection, on met en évidence un épissage alternatif accru. Cependant, les variants observés ne sont pas tous fonctionnels. Cette observation pourrait être attribuée à un fort taux d'erreur du spliceosome, ou à une réduction de la dégradation des variants d'épissage incorrects de l'ARNm. D'un point de vue évolutif, ce mécanisme pourrait réduire la conservation inter-espèce des gènes de l'immunité, favorisant ainsi la diversité du matériel génétique et des solutions alternatives pour répondre aux pathogènes [RQQ19].

2.1.2 Catégorisation des évènements d'épissage alternatif

Lorsqu'on observe *a posteriori* les transcrits produits par un gène, on peut identifier comment les exons sont utilisés dans les différents transcrits (voir par exemple une représentation des transcrits du gène ENSG00000168876 chez l'humain Figure 2.6). Par exemple, certains exons sont observés dans tous les transcrits isoformes du gène : on les qualifie d'exons constitutifs.

On a pour habitude de qualifier les évènements d'épissage alternatif suivant la manière dont on observe l'utilisation des exons : la rétention d'introns, les exons mutuellement exclusifs, les exons cassettes (ou saut d'exon), l'épissage alternatif en 5' et 3'. La description des transcrits se base sur la comparaison des transcrits deux à deux, tel qu'illustré dans la figure 2.7.

La rétention d'intron (IR) est le nom donné au cas d'un épissage entre deux exons qui n'est pas effectué. La séquence intronique normalement excisée du préARN reste alors dans l'ARNm. L'apparition d'un IR peut être dû à une mauvaise reconnaissance des sites d'épissages ou à un problème de liaison de certains inhibiteurs [Mon+19]. Même si la plupart des IR sont liés à des pathologies comme le cancer, quelques exemples d'IR fonctionnel on été découverts chez l'humain, les plantes et les champignons [Gre+20 ; Mon+19], ce qui porte à croire qu'il reste beaucoup à découvrir sur leurs fonctions. L'exclusivité mutuelles entre exons est un évenement de changement d'exon entre deux exons, qui ne cohabitent jamais sur le même isoforme. Ce changement dans la séquence d'exon à tendance à affecter un domaine protéique. Ainsi, la fonction de la protéine va être altérée [Lam+21]. L'exon cassette est l'évenement le plus couramment observé d'après la littérature [Pen+08 ; Wan+08]. Il est défini par l'exclusion d'un des exons d'un isoforme. Cet exon en moins peut affecter la fonction de la protéine, si l'exon faisait partie du domaine actif. Il peut aussi modifier le cadre de lecture du CDS et causer des *frameshift*. C'est le cas de la très étudiée myopathie de Duchenne [Oku+20].

Le choix alternatif de site d'épissage est le second événement le plus fréquent [Bra+12]. Ils sont issus d'un choix alternatif du site donneur ou accepteur qui est reconnu par le

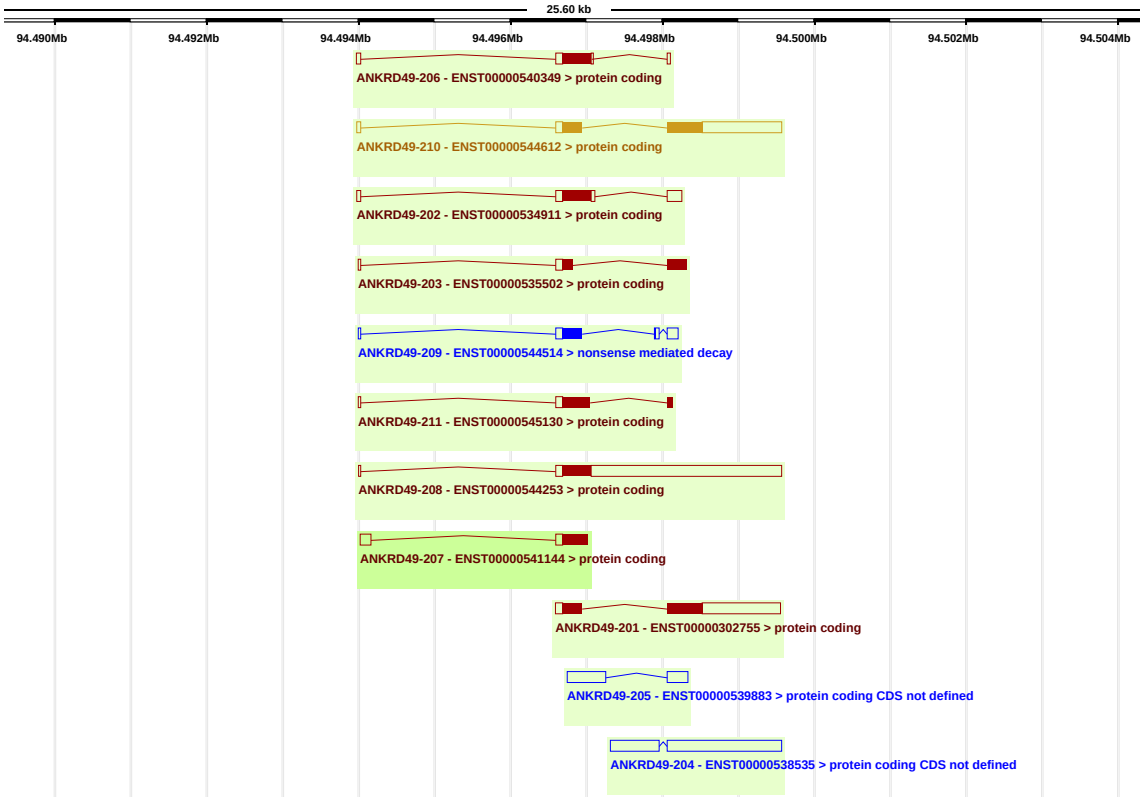


FIGURE 2.6 – Vue Ensembl des transcrits annotés CCDS du gène ANKRD49 chez l'humain. Les transcrits jaunes et rouge codent pour des protéines alors que les transcripts bleus sont des transcripts ne comportant pas d'orf ou non viable. Le transcrit jaune provient de la fusion de la base Havana avec ENSEMBL, les autres sont natifs de ENSEMBL. Ce gène comporte 11 transcrits variants d'épissage (ou isoforme).

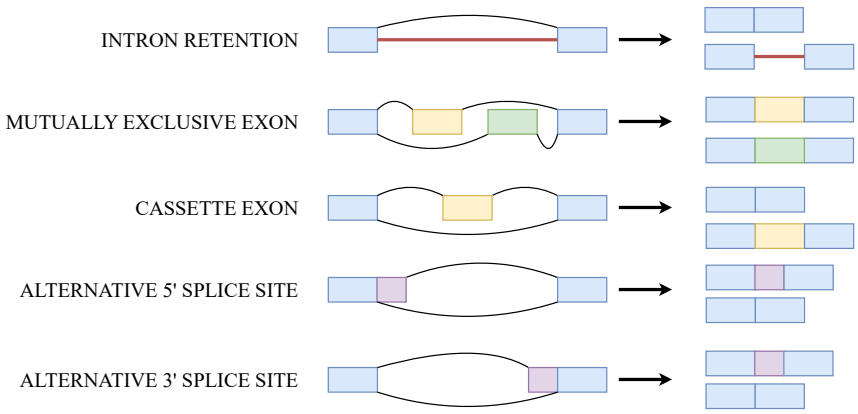


FIGURE 2.7 – Schéma des différents événements d'épissage alternatif.

spliceosome. Ce choix est régulé par les co-facteurs (U1 et *cis regulatory*, cf figure 2.5) et il peut y avoir compétition entre différent site [BG17 ; Tan+21]. Un cas particulier de ce phénomène sont les site d'épissage alternatif en tandem (TASS). Ils correspondent à une variation faible du site d'épissage et possède un motif NAGNAGs ou GYNNGYs. Ces motifs sont étudiés car ils conservent le cadre de lecture et n'engendrent pas de codons stop prématurés. A l'inverse, les évènements ajoutant ou enlevant par épissage une portion non multiple de 3 va causer un *frameshift* Cependant, toutes les événements de *frameshift* n'induisent pas une dégradation de l'ARNm ou de la protéine produite. Des travaux chez l'Homme et la souris ont mis en évidence que des événements de *frameshift* active la traduction de partie 3' UTR d'ARNms variant d'épissage [Pre+20]. Les protéines identifiées sont notamment impliqués dans des voies de signalisation. Ce mécanisme est très répandu et conservé chez les mammifères. Pour un gène, toutes les associations d'événements ne donnent pas suite à un transcrit viable. Par exemple, certains exons mutuellement exclusifs le sont du fait du décalage de phase (*frameshift*) lorsqu'ils sont tous les deux présents. Du fait de ces règles inter-exons, toutes les formes de splicing ne sont pas forcément observées. Elles peuvent être absentes soit car le spliceosome ne les crée pas, ou en petites quantités, soit car ces isoformes sont reconnus comme des ARN non-sens par le complexe EJC puis dégradés.

Erreur d'épissage La maturation des transcrits primaires par le splicesome n'est pas sans erreur. Il est possible que certains événements échappent à la régulation. Si le spliceosome possédait un taux d'erreur de seulement 0,1%, ces événements conduiraient à la production de plus de 4000 variants d'épissage différents. L'une des grandes difficultés de l'étude des variants d'épissage est de connaître quelles versions d'un transcrit sont effectivement fonctionnelles ou sont des artefacts d'erreur d'épissage [ML02]. Ainsi lors de l'analyse des erreurs de la machinerie d'épissage, la distinction entre les événements régulés ou stochastiques est très difficile. Cependant la plupart de ces erreurs d'épissage sont silencieuses pour la cellule car elles engendrent des transcrits non fonctionnels. Ces ARNs peuvent faire l'objet de décalage du cadre de lecture car leur séquence n'est pas multiple trois. Ils sont alors reconnus par la cellule via le complexe EJC (exon junction complex) puis dégradés par le complexe NMD (non-sense mediated RNA degradation complex).

2.2 Technologies de séquençage d'ARN

Les technologies de séquençage sont l'ensemble des techniques qui permettent de reconnaître la suite d'acides nucléotidiques formant une molécule d'ADN ou d'ARN. L'étude de cette séquence donne accès à une information capitale pour comprendre le vivant. En séquençant une molécule d'ADN, il est possible d'assembler un génome qui permet d'établir une référence pour une nouvelle espèce, d'identifier des variants chez une espèce connue ou encore d'étudier l'évolution inter-espèces. Tout comme on peut séquencer un génome, il est possible de séquencer les transcrits. Le séquençage des transcrits permet d'étudier l'expression des gènes *in situ*. Il est alors possible de quantifier cette expression mais aussi de la caractériser, par exemple en identifiant des variants de séquence ou d'épissage.

Pour toutes ces analyses, différentes technologies sont disponibles. La plus ancienne, la technologie Sanger (1951) a permis de séquencer les premiers génomes. Les plus récentes, la technologie de séquençage de deuxième (années 2000) et troisième génération (années 2010), respectivement séquence courte lecture et longue lecture ont permet d'accélérer le débit et diminuer les coûts. Ainsi le séquençage du premiers génome humain, réalisé dans le cadre du Human Genome Project [Lan+01] a pris 13 ans et a coûté 3 milliards d'euros alors qu'aujourd'hui une expérience nécessitant le séquençage du génome entier prend quelques jours et coûte environ 7000 euros (culture des cellules comprises) [Sch+20]. On discute ici des avantages et inconvénients des techniques de séquençage de deuxième et de troisième génération pour l'identification des variants d'épissage.

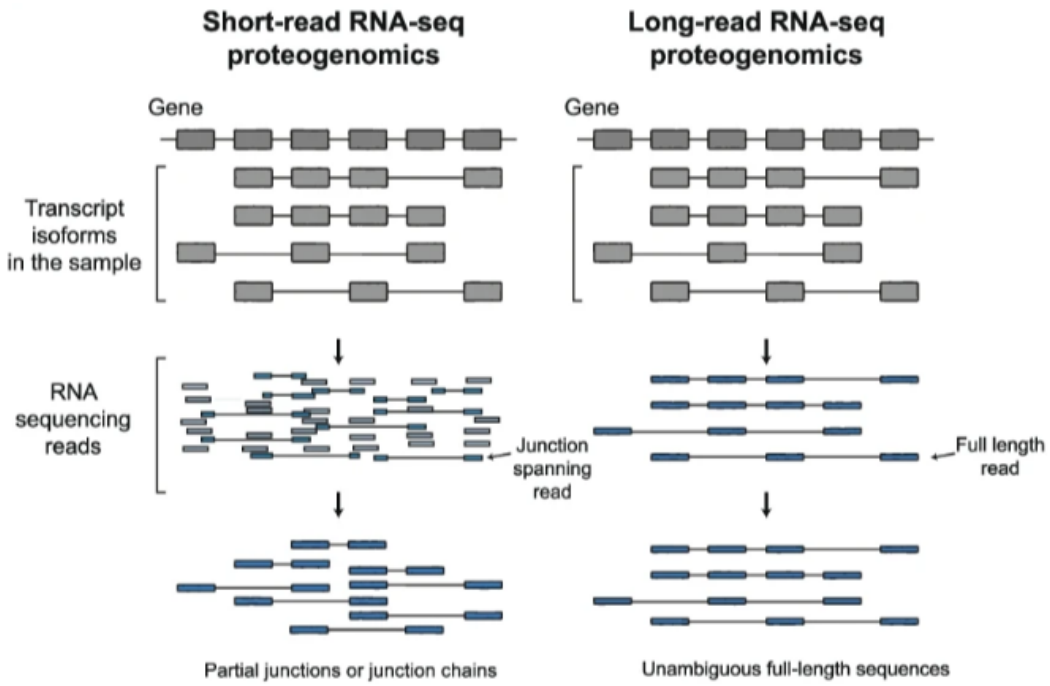


FIGURE 2.8 – Figure de comparaison des caractéristiques de séquençage des lectures courtes contre lectures longues pour la détermination des isoformes alternatifs. A gauche de l'image, l'ensemble des lectures courtes séquences soutiennent des jonctions exoniques contenus dans l'échantillon. Cependant, il est impossible par reconstruction de retrouver l'information compositionnelle initiale. A l'inverse à droite, les lectures longues couvrent l'entièreté des isoformes et permettre de résoudre les isoformes exprimés sans ambiguïté. Tirée de [Mil+22].

2.2.1 Technologies de séquençage pour l'étude de l'épissage alternatif

Les technologies de séquençage de deuxième et de troisième génération produisent des données aux caractéristiques différentes. Celles-ci doivent être prises en compte pour l'étude de l'épissage alternatif.

La technologie de séquençage seconde génération (Illumina) produit des lectures courtes avec un taux d'erreur très faible $\tilde{0}.01\%$ de type substitution. Cependant, la longueur de ces lectures limitée à 300 paires de bases ne permet pas de séquencer un transcrit dans sa pleine longueur.

Pour pallier ce problème, différents outils d'analyse passent par une étape d'assemblage du transcrit à partir des lectures courtes (illustration Figure 2.8) . Cette étape repose sur des validations statistiques pour prédire les isoformes ayant les suites de jonctions d'épissage les plus vraisemblables [Nip+20 ; Bus+19]. Ainsi, les prédictions ne se basent pas sur des observations directes des transcripts isoformes mais sur la reconstruction d'un chemin de jonctions d'épissage détectés par les lectures courtes. Cela peut poser des problèmes d'ambiguïté dans le cas où il y a plusieurs évènements de saut d'exon détectés, les lectures courtes ne pourront pas résoudre la combinaison entière des exons de chaque isoforme ou encore s'il s'agit d'un simple ou double saut dans certains cas.

À l'inverse, les technologies de séquençage de troisième génération, produisent des lectures longues mais avec un taux d'erreur beaucoup plus important ($\tilde{5}$-10%) majoritairement de type insertion délétion [Ses+19]. La longueur de séquençage possible en fait par contre un outil très intéressant pour l'étude du transcriptome puisque les lectures peuvent atteindre une longueur de 10kbp. Il est donc complètement possible de séquencer des transcrits en pleine longueur et ainsi de s'affranchir de l'étape d'assemblage (illustration Figure 2.8). Comparativement aux technologies de séquençage de $2^{\text{ème}}$ génération, le séquençage pleine longueur permet de faire des observations directes des isoformes présents dans l'échantillon

sans passer par une phase d'assemblage qui ne fournit que des prédictions. Les ambiguïtés sont ainsi levées. La figure 2.8 présente un cas de non résolution de la séquence d'exons par le séquençage court, alors que le séquençage pleine longueurs permet de le résoudre.

Les technologies de séquençage de troisième génération sont donc fortement préconisées pour l'étude de l'épissage alternatif et ont été identifiées comme tel depuis leur mise en place [Byr+19]. Néanmoins il subsiste des difficultés que nous détaillerons dans la section suivante. Enfin, et bien que dans le cadre de ce travail nous ne nous intéressions pas aux méthodes hybrides, il est à noter que la complémentarité des deux technologies de séquençage peut être bénéfique, les lectures courtes permettant de pallier le taux d'erreur des lectures courtes.

2.2.2 Caractéristiques des différentes technologies de séquençage 3ème génération

Le séquençage longue lecture des molécules d'ADN et de ARN a connu un développement fort dans les dix dernières années. Il existe deux acteurs incontournables sur le segment des technologies de séquençage longue lecture : Oxford Nanopore et Pacific Bioscience.

Séquençage Nanopore

La technologie de séquençage par Nanopore repose sur l'utilisation d'un pore protéique de taille nanométrique, relié à une membrane de polymères isolants. L'activité de ce pore fournit des informations sur la séquence de l'acide nucléique qui le traverse. Le passage de l'acide nucléique à travers le pore est contrôlé par un moteur protéique, qui fait avancer la molécule nucléotide par nucléotide. Les micro-changements du potentiel électrique entre les faces *cis* et *trans* de la membrane sont caractéristiques de chaque nucléotide, ce qui permet de déduire la séquence de l'acide nucléique [Wan+21]. Cette technologie permet le séquençage de molécules extrêmement longues ($>$ 4mb). Oxford Nanopore offre une alternative économique et portable grâce à son Minion.

Depuis 2014, des améliorations continues du pore et du moteur protéique, éléments clés du processus, ont été réalisées. Elles ont permis de passer le taux d'erreur de séquençage de près de 15% à 5% de taux d'erreur. Sept versions de cette technologie ont été développées en six ans : R6 (juin 2014), R7 (juillet 2014), R7.3 (octobre 2014), R9 (mai 2016), R9.4 (octobre 2016), R9.5 (mai 2017), R10 (mars 2019) et R10.3 (janvier 2020). Ces mises à jour ont progressivement amélioré la précision de la reconnaissance des bases, la taille des lectures (reads) et le rendement du séquençage, notamment avec l'arrivée du PromethION, qui a augmenté la profondeur de séquençage. La diminution du taux d'erreur de séquençage est également attribuée à l'amélioration des logiciels de traitement du signal électrique en sortie. Différentes méthodes de reconnaissance du signal se sont succédées : basées sur les modèles HMM et les réseaux de neurones en 2016, sur les données brutes en 2017, utilisant un modèle flip-flop en 2018 et des algorithmes d'entraînement spécifiques depuis 2019 [Wan+21]. Nanopore propose deux types de protocole de séquençage : le protocole cDNA et le RNA-direct. La différence majeure de ces deux techniques réside dans le fait que dans la première les ARNm seront rétro-transcrits en ADN avant le séquençage, alors que dans la seconde, ce sont bien les ARNm bruts qui passeront dans le pore de séquençage. L'intérêt du RNA-direct est donc d'éviter l'étape de rétro-transcription et les biais de séquences qu'il peut engendrer. Il permet une résolution et une quantification plus fiables des séquences des ARNm dans l'échantillon [Ses+19].

Séquençage PacBio (SMRT)

La technologie de séquençage PacBio utilise leur technologie propriétaire SMRT (Single Molecule Real-Time) pour déterminer la séquence d'une molécule d'ADN. Les molécules

d'acides nucléiques sont encapsulées dans des structures appelées "SMRTbell", des molécules d'ADN simple brin repliées sur elles-mêmes. Ces SMRTbell sont séquencées dans des "Zero-Mode Waveguide" (ZMW), qui offrent un environnement optimal pour la détection lumineuse. Au fond de chaque puits ZMW, une ADN-polymérase unique est fixée. La séquence de l'ADN est obtenue en suivant l'activité de réplication de l'ADN simple brin par cette polymérase. Les nucléotides fluorescents (A, T, G, C) présents en solution émettent de la lumière lorsqu'ils sont incorporés par la polymérase. Le séquençage est donc réalisé en lisant le film des flashs lumineux produits lors de l'incorporation des nucléotides. Parfois, le même brin d'ADN peut être séquencé plusieurs fois. Dans ce cas, la séquence finale est déterminée par une méthode de consensus circulaire (Circular Consensus Sequencing, CCS), où les séquences multiples sont combinées pour obtenir une séquence consensus. Cette méthode est très performante, permettant à la technologie PacBio de générer des lectures avec un taux d'erreur comparable à celui des lectures courts de la technologie Illumina [RA15 ; Kan+21].

Challenges du RNA-seq 3ème génération

Bien que prometteuses, les technologies RNA-seq 3ème génération posent un certain nombre de challenges [Byr+19] :

- intégrité des molécules d'ARN séquencées : les méthodes d'extraction doivent garantir d'obtenir l'ARN complet ;
- biais de longueur : l'amplification ainsi que la technologie de séquençage elle-même ont tendance à privilégier les transcrits courts, ce qui rend complexe l'identification de transcrits longs (de longueur supérieure à 2 kb), même si le biais dû à l'amplification pourrait être surmonté grâce au séquençage direct ;
- débit : pour explorer toute la diversité des transcriptomes de mammifères, il faudrait envisager de l'ordre de 100 millions de lectures pleine longueur par tissu ou organe ;
- qualité des lectures : bien que la technologie ait progressé ces dernières années, le taux d'erreur reste important, la question de l'utilisation de méthodes de correction des lectures se pose.

Comparaison des technologies

Dans [Ses+19], les auteurs comparent les données de séquençage obtenues avec les technologies Oxford Nanopore (cDNA et Direct RNA) et Illumina (cDNA). Cette étude révèle que l'alignement des variants d'épissage alternatif varie en fonction des outils, des technologies et des protocoles de séquençage utilisés. Pour évaluer la capacité des technologies à séquencer correctement. Les résultats de l'étude montrent que le taux de jonctions exoniques GT-AG sont de 98,5 % pour des lectures parfaites provenant de gènes connus de la souris, en comparaison de 80,7 % pour les lectures longues Nanopore RNA Direct et 67,7% pour le cDNA alignés sur ces gènes. Pour les lectures longues de PacBio, le meilleur taux de GT-AG introns atteint 96,4 %. Cela indique que les aligneurs lectures longues ont encore une marge de progression En moyenne, pour le séquençage longues lectures, la couverture des lectures est inférieure à 80 % de leur longueur totale majoritairement allouée à la dégradation de l'ARNm.

Biais de Séquences

Un des biais majeurs du séquençage lié à la technologie est la présence de séquences poly(T) internes dans les lectures. Ces séquences polyT peuvent contaminer les lectures, posant des défis pour déterminer si une séquence contenant des polyT internes est correctement séquencée. Ce biais est plus prononcé pour les séquences polyT que pour les séquences polyA, probablement en raison de la potentialisation des chaînes polyT pendant la première synthèse et de la troncature au niveau des queues polyT. La méthode cDNA a tendance à surexprimer les lectures tronquées [Ses+19].

Coût d'utilisation et profil d'erreur.

TABLE 2.2 – Comparaison du coût et des caractéristique des technologies Illumina, ONT Nanopore et Pacific Biosciences pour le séquençage. Donnée tiré de [SN21], [Liu+24], [LVE20].

Séquenceur	Coût par Gigabase (Gb)	Longeur de séquençage	Précision de séquençage et type d'erreur
Illumina NextSeq 550	\$40-63	75 à 150 bp	>$\tilde{9}$9.6% & Substitutions de bases, rares insertions et délétions
Illumina NovaSeq 6000	\$10-35	50 à 300 bp	99.9% ,
ONT MinION/GridION	\$50-2000	10 000 à 30 000 bp jusqu'à 2.3 millions bp	87-98% erreur de type insertions, délétions
ONT PromethION	\$21-42	10 000 à 100 000 bp et lectures ultra-longues	-
PacBio Sequel II	\$70-100	10 000 à 60 000 bp	>99% (lectures HiFi) & Insertions, délétions, substitution

Le choix d'une plateforme de séquençage dépend des objectifs spécifiques du projet et des ressources disponibles. Le tableau comparatif 2.2 illustre les différences en termes de coût, de longueur de séquençage et de précision entre les séquenceurs d'Illumina, Oxford Nanopore Technologies (ONT), et PacBio. Pour des projets nécessitant l'analyse de structures complexes comme les isoformes d'épissage, ou pour des applications où la longueur des lectures est critique, par exemple, dans l'étude de régions génomiques hautement répétitives ou dans le séquençage *de novo*, les technologies de lecture longue d'ONT et PacBio sont particulièrement avantageuses. Elles permettent d'obtenir une vue plus complète et moins fragmentée des génomes et transcriptomes. Elles restent donc intéressantes malgré un coût plus élevé et un taux d'erreur plus important en comparaison à Illumina.En revanche, pour des applications telles que le "SNP calling" ou la validation de jonctions d'épissage, où la précision des lectures est primordiale, le séquençage Illumina, avec leur coût réduit et leur haute précision, représentent une option plus appropriée. Leurs lectures courtes mais extrêmement précises sont idéales pour des analyses détaillées de variations génétiques à petite échelle.

2.3 Méthodes et outils pour l'analyse de données RNA-seq

Dans cette partie, nous allons discuter de méthodes et d'outils pour l'analyse de données RNA-seq avec des lectures longues. Nous débutons par une présentation des outils permettant l'alignement et la correction des lectures, dont nous discuterons les effets sections 4.3 et 4.3.1. Puis nous présenterons un panorama des outils permettant d'identifier des isoformes à partir d'un jeu de données RNA-seq lectures longues.

2.3.1 Méthodes d'alignements *splicé*

exonerate : modèle est2genome Les séquences des reads sélectionnées à l'étape précédente sont alignées grâce au programme exonerate [SB05]. Cet outil, basé sur une stratégie de type "seed and extend", produit un alignement composé de High-scoring Segment Pairs (HSPs). La programmation dynamique utilisée ici suit le modèle "EST to genome", basé sur l'hypothèse selon laquelle un ADNc est aligné sur une séquence génomique. Les HSPs produits sont donc attendus comme devant correspondre aux exons du read aligné.La stratégie d'alignement splicé assure un alignement de toute la longueur du read sur le gène d'intérêt. C'est un atout car du point de vue biologique toute la séquence doit être alignée. L'aligneur utilise une programmation dynamique de type *bounded sparse*. On choisit de renvoyer seulement l'alignement de meilleur score pour chaque read sur la référence.

minimap2 minimap2 [Li18] est le mappeur à l'état de l'art pour l'alignement des lectures longues contre une référence. Il fonctionne en deux étape : une indexation du génome, qui est fait hors du process d'alignement, puis il utilise cet index pour trouver rapidement la position des minimizers des lectures sur ce derniers et les organiser en chaines[Sad+23]. L'indexation crée un index réutilisable du génome. Cet index stocke les minimizers des

k-mères dans un multimap grâce à une table de hashage. Les minimizers sont les clés et les valeurs sont les positions sur le génome. Le chainage est rapide et identifie des correspondances exactes, courtes de longueur fixe (chainage de minimizers) entre une lecture et une séquence de référence. Lorsque minimap2 traite une lecture, les minimizers de la lecture sont utilisés pour trouver des ancres, en interrogeant l'index du génome de référence pour des « match » exactes. Ces ancres sont ensuite triées en fonction de leur position dans la référence et sont sélectionnés pour le chaînage. Le chaînage prend les ancres triées en entrée et identifie des chaines ordonnées et colinéaires d'ancres, de sorte qu'aucune ancre n'est utilisée dans plus d'une chaîne. Le chainage est fait par programmation dynamique. Il sélectionne un sous-ensemble de chaînes alignées sur la référence cible. Afin d'aligner les bases restantes inter-ancres, il utilise une programmation dynamique Needleman-Wunsch [NW70] avec la formulation Suzuki-Kazahara [SK18]. Il est important de noter que par défaut avec l'option 'alignement splicé', l'alignement inter-ancre va préfèrer GT[A/G]..[C/T]AG sur GT[C/T]..[A/G]AG, puis sur les autres signaux d'épissage. Cet heuristique permet des bonnes performances d'alignement sur lectures bruitées comme Nanopore, mais il est conseillé de les désactiver pour effectuer l'alignement sur des lectures de bonnes qualités comme PacBio.

2.3.2 Méthodes de correction des lectures longues

La qualité de séquençage est une des limites fortes pour l'utilisation des lectures longues. Dans le cadre de l'identification des isoformes alternatifs, les erreurs de séquençage peuvent altérer l'alignement et donc la bonne détection des jonctions exoniques. Pour pallier ce problème, des méthodes de correction des séquences des lectures ont été développées. Il existe deux types de stratégies : l'auto-correction des lectures (lectures longues seules) et la correction hybride (lectures longues et courtes). Ici, nous ne discutons que de l'auto-correction, la méthode que nous proposons étant basée uniquement sur des données de lectures longues.

isONcorrect [Sah+21] est conçu pour corriger les erreurs de séquençage des lectures longues, sans apport externe, spécifiquement dans le cadre de la transcriptomique. La méthode est basée sur le clustering de lectures puis la correction indépendante de chaque cluster. L'outil isONclust [SM20] est utilisé pour le clustering. Le clustering permet de rassembler les séquences par gène ou famille de gènes plutôt que par isoforme. Cela permet d'augmenter le nombre de séquences à corriger dans le batch et donc d'améliorer le support de chaque position. Dans chaque lecture, une liste d'ancres est établie (en utilisant des *minimizers*). Le but de ces ancres est d'identifier et d'organiser des zones similaires entre les lectures. Les intervalles entre chaque ancre sont calculés pour chaque lecture, puis ils sont pondérés en fonction de leur fréquence d'apparition dans l'ensemble du cluster de lectures. Cette méthode permet de distinguer les motifs communs entre les lectures des erreurs de séquençage. Les intervalles sont sélectionnés de manière à couvrir la plus grande partie possible de chaque lecture tout en maximisant le poids représentatif de chaque intervalle. Les intervalles équivalents de chaque lecture sont ensuite alignés par SPOA [Vas+17], un outil qui permet de générer une séquence consensus. L'algorithme prend également en compte les éléments répétés et les queues PolyA en masquant les ancres contenant ces motifs. Cette méthode assure une amélioration significative de la qualité des données de séquençage à longues lectures sans besoin d'apport d'information supplémentaire.

Dans le cadre de l'identification d'isoformes alternatifs, une auto-correction peut avoir un effet délétère sur l'identification de certains variants qui pourraient être gommés par l'établissement d'une séquence consensus. Ces problèmes potentiels sont évoqués dans le chapitre 4 où nous discutons de l'apport de la correction aux prédictions.

TABLE 2.3 – Récapitulatif et comparaison des outils d'identification d'isoformes alternatifs d'épissage sortis depuis 2020.

Outil	Année de publication	Guidé par une séquence de référence	Guidé par des annotations	Applique une correction des jonctions d'épissage	Applique une correction des lectures	Type de lectures
FLAIR	2020	Génome	Oui/Non	Oui/Non	Non	longues
StringTie2	2020	Génome	Oui/Non	Oui	Oui/Non	longues+courtes
Freddie	2023	Génome	Non	Oui	Non	longues
TAMA	2023	Génome	Non	Oui	Non	longues
UNAGI	2023	Génome	Non	Oui	Non	longues
Bambu	2023	Génome	Oui/Non	Oui	Non	longues
IsoQuant	2023	Génome	Oui/Non	Oui	Non	longues
Mandalorian	2023	Génome	Oui	Oui	Non	longues
isONform	2023	Sans référence	Non	Oui	Non	longues
Espresso	2023	Génome	Oui/Non	Oui	Non	longues

2.3.3 Méthodes d'identification des isoformes issus de l'épissage alternatif à partir de lectures longues

La question de l'identification des isoformes d'épissage à partir de données de séquençage longues lectures a connu un engouement fort depuis 2020 avec un pic à l'été 2023 (le tableau 2.3 répertorie les années de publication). Ce gain en intérêt est porté par le développement et l'amélioration continue des technologies de séquençage longues lectures. Elles rendent possible l'analyse de plus en plus précise des variants d'épissages et la quantification de ses transcrits. Ces méthodes se distinguent par leur diversité d'approche et de choix méthodologiques pour traiter les difficultés inhérentes au séquençage de $3^{\text{ème}}$ génération listées plus haut. De manière plus profonde, toutes n'adressent pas tout à fait le même problème : possibilité ou non d'utiliser des annotations pour identifier les transcrits ; possibilité ou non de faire de la correction, ou correction hybride ; possibilité ou non d'utiliser une séquence de référence. Certains ajoutent également une option de quantification. Le tableau 2.3 donne un résumé des différentes options.

Classification selon le besoin de guidage

La diversité des outils reflètent la diversité des usages lié à l'identification d'isoformes d'épissage 2.3. La plupart des outils propose d'utiliser une annotation de référence pour guider l'identification des isoformes. Cette approche est pensée pour identifier le plus d'isoformes correspondant à des données connues. Elle peut être pertinente pour du diagnostic. Cependant, l'utilisation d'annotation est à double tranchant. Elle augmente le risque de détection d'isoformes :

— faux positifs connus
— faux négatifs non connus

De ce point de vue, le développement d'outil ne nécessitant pas d'annotation de référence permet d'obtenir des résultats non biaisés. Certains outils construisent leur méthode sans utiliser d'annotation comme Freddie, TAMA, UNAGI et RNA-Tailor. La contrainte de ces méthodes est de trouver sans a priori un bon compromis entre sensibilité du signal et filtrage du bruit issu de l'alignement et du séquençage. Il est tentant de varier les usages c'est pourquoi certaines intègrent à la fois des modules guidés et non guidés, comme StringTie2, FLAIR, IsoQuant, Bambu, ESPRESSO.

Approches et stratégies algorithmiques de différents outils

StringTie2 [Kov+19] propose une approche hybride lecture courte et lecture longue. Sa méthode tire partie de la complémentarité entre les deux technologie, l'une possède un faible taux d'erreur et la seconde propose un séquençage pleine longueur des transcrits.

Dans un premiers temps, l'alignement des lectures longues permet de créer un graphe d'épissage. Cependant les erreurs de séquençage des lectures longues créent des incertitudes en multipliant les noeuds de part et d'autres des jonctions d'épissages. Pour cela, Stringtie impose un nombre maximale de noeud (par defaut 1000) et supprime ou fusionne les noeuds en commençant par les plus faiblement supportés jusqu'à atteindre le seuil de noeud. La correction des alignements splicés se fait avec la fusion des noeuds. Elle est réalisée dans une fenêtre de 10 bp du noeud le moins supporté vers le noeud le plus supporté. Les courtes lectures sont assemblées en « *super-reads* » par l'assembleur MaSuRCA [Zim+13]. Les « *super-reads* » et les lectures non assemblées sont alignées sur le génome. Leur alignement permet de valider les jonctions d'épissage, à savoir que les lectures courtes ayant plus de 1% d'erreur sont éliminées.

IsoQuant [Prj+23] propose d'aligner les lectures longues en entrée avec minimap2, mais peut également prendre en entrée un fichier BAM. À partir des alignements, il construit un graphe d'introns à partir des lectures alignées, utilisant des chemins à travers le graphe pour reconstruire les isoformes de transcrits. Il dispose de plusieurs algorithmes de correction et de sélection des parties moins supportées du graphe : la correction des sites d'épissage est active si une annotation est donnée. Elle réassigne les jonctions à la jonction annotée la plus proche dans une fenêtre de 6 nt pour les lectures ONT et 4 nt pour PacBio. Le reste des jonctions faiblement supportées est traité par simplification du graphe. Un noeud peut être supprimé s'il n'est pas annoté. Il considère que les erreurs de prédiction des sites d'épissage contaminants se situent sur des petites longueurs de nucléotides. Ainsi, un noeud est supprimé s'il représente une jonction ayant deux fois moins de support de lectures qu'un noeud alternatif dans une fenêtre de 20 nt (10 pour PacBio). Il effectue également une quantification basée sur le nombre d'alignement des lectures. Le filtrage des lectures par support (un minimum de 5 pour les ONT) qui permet de classer les isoformes prédits dans les catégories : *consistent*, *ambiguous*, et *inconsistent*.

ESPRESSO [Gao+23] prend en entrée un fichier d'alignement BAM. Sa méthode commence par la détection des jonctions d'épissage solides contenu dans l'alignement : pour cela soit la jonction appartient à l'ensemble des jonctions connues données en entrée soit elle est supportée par une jonction canonique et au moins deux lectures avec un alignement sans erreur de part en d'autre de sa jonction d'épissage. Une fois que l'ensemble des jonctions solides sont identifiées, pour chaque lecture, les sites d'épissage non solides qui possèdent une jonction solide dans une fenêtre de 35 nt vont être corrigés par réalignement en utilisant blastn [Cam+09]. Le réalignement de cette sous séquence est réalisé contre la concaténation de 50 nt autour de jonction solide la plus proche et est validée si le match est sans erreur dans une fenêtre de 10 nt. Après ces corrections, les isoformes ayant l'ensemble de leurs sites solides sont exportés en GTF. Espresso fait également de la quantification des isoformes avec un algorithme d'« expectation minimization ».

Bambu Sa méthode [Che+23] propose l'identification et la quantification des transcrits à partir des données de séquençage RNA-seq à lecture longue en utilisant l'apprentissage automatique, fait à partir d'annotations, pour corriger les erreurs d'alignement et prédire les isoformes d'épissage. Elle classe les lectures en groupes basés sur les motifs d'épissage, entraîne un modèle en utilisant des caractéristiques telles que les comptages de lectures et les motifs de sites d'épissage. Le modèle les compile en un seul score appelé TPS (*transcript probability score*). Pour identifier de nouveaux isoformes, le modèle estime un « novel discovery rate ». Ce taux peut être utilisé par Bambu pour ajuster dynamiquement les annotations en fonction du contexte de l'échantillon, améliorant ainsi la précision dans la découverte et la quantification des transcrits. Il est défini comme la fraction des transcripts non annotés qui ont un TPS supérieur ou égal au transcripts annotés. Ses auteurs insistent sur l'intérêt de l'intégration des annotations ajustées au contexte de l'expérience de sé-

quençage (tissus spécifique par exemple) et proposent d'utiliser des modèles pré-entraînés pour les génomes moins annotés.

FLAIR [Tan+20] possède une méthode scindée en deux modules indépendants : correct et collapse. Le module "correct" est utilisé pour corriger les jonctions d'épissage directement dans un bed créer à partir du fichier BAM en sortie d'alignement. Elles sont corrigées de manière à coïncider avec les jonctions annotées données, dans une fenêtre de 10 nt, réduisant considérablement la variabilité dans le jeu de donnée et corrigeant les erreurs d'alignement au niveau des points de jonctions vers les jonctions connues. Le module « collapse » réaligne les lectures en entrée sur l'ensemble des séquences des isoformes épissés prédits extraits d'un fichier BED donné en entrée. Celui-ci peut provenir de l'aligneur directement ou sortir du processus de correction. Les isoformes sont extraits selon différents modes : par défaut toutes les structures introniques uniques sont conservées, mais il est aussi possible d'éliminer les structures exoniques incluses les unes dans les autres pour réduire le nombre d'isoformes séléectionnables. Les isoformes sélectionnés sont ceux qui ont reçu un support de lecture supérieur à 5 après réalignement par minimap2 des lectures longues directement sur la base de donnée des séquences des isoformes sélectionnables.

Freddie La méthode de Freddie [Ora+23] fonctionne sans annotation et se découpe en cinq étapes. Si les lectures ne sont pas déjà alignées, Freddie utilise minimap2 pour aligner les lectures longues. Ensuite les lectures sont séparées sur le génome en cluster, représentant les gènes. Puis dans chaque cluster, une courbe de couverture en lecture est générée, puis un filtre de Gauss est appliqué sur cette dernière pour identifier les points de jonctions vraisemblables. La sélection de ces « points de segmentation » est ensuite soumis à la sélection par une fonction de *scoring* maximisant la couverture. Enfin les isoformes prédits sont sélectionnés par l'algorithme MErCi (Minimum Error Clustering into Isoforms), formulé comme un « *integer linear programming (ILP) problem* ». Il permet d'optimiser l'assignation des lectures dans les isoformes similaires suivant leur couverture. Durant chaque itération de ce processus, une fonction de score évalue si deux isoformes sont suffisamment similaires, ou doivent subir des corrections, et peuvent être placés dans la même « bin ». Le processus continue tant qu'il reste des isoformes à placer dans une « bin ».

isONform L'algorithme isONform [PS23] détermine l'ensemble des isoformes à partir de données de séquençage à longues lectures, indépendamment des génomes de référence ou des annotations. Cela le rend particulièrement adapté à l'utilisation sur des espèces non-modèles. Les lectures sont groupées en intervalles en utilisant des paires de minimizers pour ancres et leur assignation assistée par le « Weighted Interval Scheduling » qui maximise le poids de paires de minimizers. Il traite les lectures « clusterisées » et corrigées (si demandé) par isONcorrect [Sah+21]) pour construire un graphe acyclique dirigé en utilisant les paires de minimizers non chevauchantes du cluster de lectures en tant que noeuds du graphe. Le graphe subit une simplification grâce à un algorithme itératif qui élimine les noeuds peu représentés et les erreurs de séquençage tout en exerçant un compromis pour préserver les différences cruciales entre exons. Les chemins de ce graphe simplifié représentent des isoformes distincts. Une fois les isoformes identifiés, une séquence consensus de cet isoforme est produite à partir d'un alignement multiple des lectures par SPOA [Vas+17]. Un support de lectures de 5 est nécessaire pour qu'un isoforme soit généré.

Conclusion

L'épissage alternatif est un mécanisme clé permettant de générer une grande diversité de protéines à partir d'un nombre limité de gènes. Ce processus se produit après la transcription de l'ADN en ARN pré-messager, où les introns non codants sont retirés et les exons codants sont reliés pour former l'ARN messager mature. L'épissage alternatif peut aboutir

à la production de différents isoformes de protéines en incluant ou excluant certains exons, influençant ainsi les fonctions des protéines produites. Ce mécanisme est crucial dans divers processus biologiques et est impliqué dans la régulation de nombreuses fonctions cellulaires. Son étude a été permise par le développement du séquençage des acides nucléotidiques. Les technologies de séquençage d'ARN ont considérablement évolué, passant des méthodes Sanger au séquençage de deuxième génération, Illumina, puis aux technologies de troisième génération, telles qu'Oxford Nanopore et PacBio. Ces dernières permettent de séquencer des lectures longues, offrant une meilleure capacité à détecter les isoformes d'épissage et les structures complexes des ARN. Cependant, chaque technologie présente des avantages et des inconvénients qui vont répondre des besoins différents. L'analyse des données RNA-seq nécessite des outils pour l'alignement et la correction des lectures. Les méthodes d'alignement splicé, comme Exonerate et Minimap2, sont essentielles pour cartographier les lectures sur le génome de référence. Les corrections des lectures longues sont cruciales pour minimiser les erreurs de séquençage et améliorer la précision de l'identification des jonctions d'épissage. Des outils comme isONcorrect sont utilisés pour ces corrections. En outre, l'identification des isoformes d'épissage à partir des données RNA-seq peut être réalisée à l'aide d'outils spécifiques, chacun ayant ses propres avantages en fonction de l'utilisation de séquences de référence ou d'annotations .

Chapitre 3

RNA-tailor

Sommaire

Introduction

Nous décrivons dans ce chapitre la méthode que nous proposons, RNA-tailor. Elle a pour objectif d'identifier, pour un gène d'intérêt dont on dispose d'une séquence de référence, et une expérience de transcriptomique de lectures longues, les transcrits isoformes alternatifs d'épissage sans connaissance préalable des annotations.

Le point de vue original de RNA-tailor est de se placer à une échelle différente par rapport aux outils existants, à savoir celle du gène et non celle du génome. Ce positionnement donne sa singularité à RNA-tailor, il permet de donner un rendu détaillé de la structure des isoformes et de l'alignement pour chacune des lectures grâce à un export XLSX. Il détaille les alignements de chaque lecture sur un gène. Travailler au niveau du gène offre aussi la possibilité d'explorer finement les résultats, voire de comparer les résultats en fonction du choix des paramètres de la méthode. RNA-tailor se différencie également par le choix d'un aligneur « splicé », exonerate, pour fournir les alignements sur lequel la détection d'isoforme se fera. Enfin le pipeline incorpore également un outil d'auto-correction des lectures, isONcorrect, qui permet d'améliorer la qualité de séquençage des lectures.

L'outil est disponible sur le dépôt git : `https://gitlab.univ-lille.fr/bilille/RNA-tailor`.

Notre travail a également amené à une réflexion à propos de la classification des isoformes prédits et sur leur capacité à produire une protéine.

3.1 Présentation générale

RNA-tailor se présente sous la forme d'un pipeline d'analyse de l'ensemble des lectures longues qu'il prend en entrée. Une vue schématique est donnée Figure **??**.

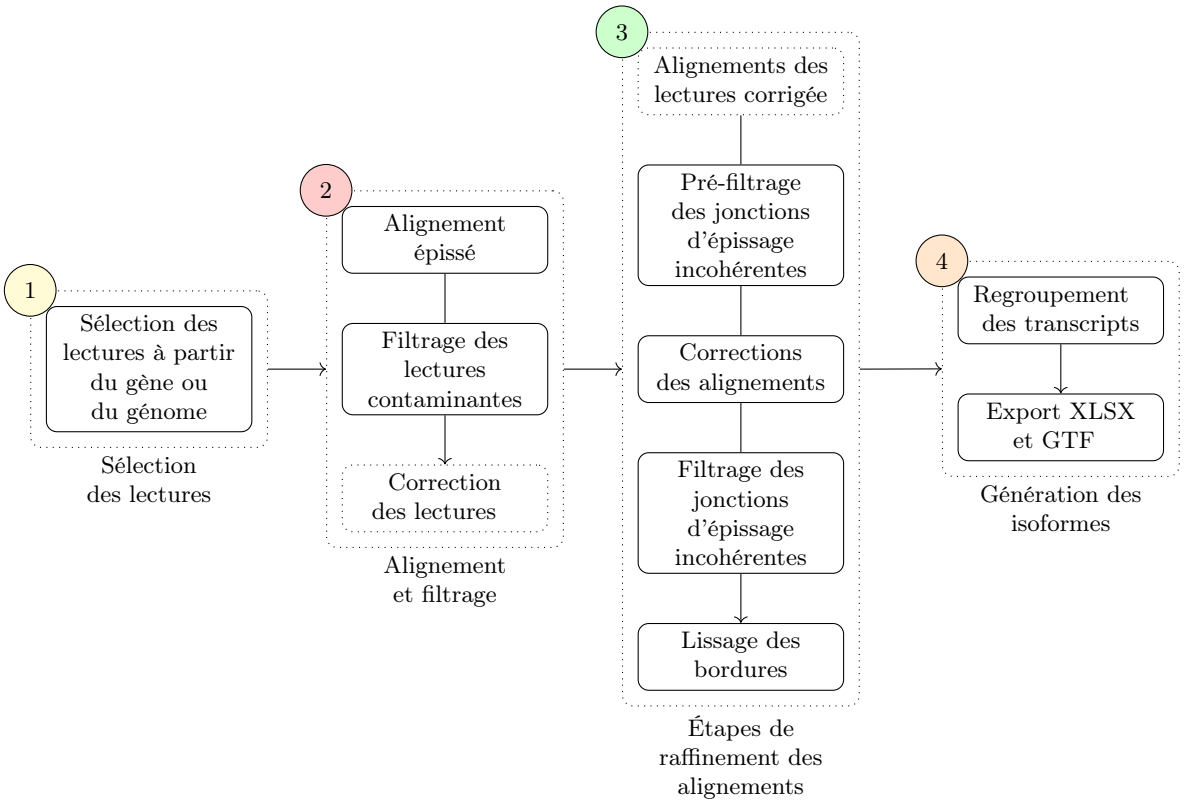


FIGURE 3.1 – Vue d'ensemble du pipeline d'analyse de RNA-tailor.

La première étape concerne la sélection des lectures d'intérêt à partir desquelles commencer l'analyse. Cette étape prend en entrée l'ensemble des lectures et la séquence de référence (gène ou génome et locus). Elle produit un sous-ensemble des lectures. La seconde étape a pour but de filtrer et de raffiner ce sous-ensemble de lectures. On va étudier l'homogénéité des lectures entre-elles et appliquer un algorithme de correction à celles-ci. On obtient un ensemble de lectures homogènes et corrigées que l'on aligne sur la séquence de référence du gène d'intérêt avec un outil d'alignement prenant en compte l'épissage. En sortie on a donc un alignement, qu'on qualifiera d'alignement *splicé*, de chaque lecture sur la séquence de référence du gène d'intérêt. La troisième étape est le cœur de RNA-tailor. Il s'agit de corriger les résultats d'alignement autour des jonctions d'épissage. Dans un premier temps on détecte les jonctions d'épissage incohérentes et on filtre à nouveau les lectures, puis on applique diverses étapes de corrections consistant à réaligner ou à corriger les alignements. En sortie nous disposons pour chaque lecture d'une structure introns/exons. La quatrième et dernière étape consiste à regrouper les lectures ayant la même structure intronique pour obtenir les isoformes.

3.2 Choix stratégiques

Lors du développement de RNA-tailor nous avons réalisé certains choix concernant les deux premières étapes du pipeline que nous discutons ici.

3.2.1 Sélectionneur de lectures.

La première étape du pipeline consiste en la sélection, dans le jeu de données, des lectures susceptibles d'être un transcrit issu de notre gène d'intérêt. Le but est de les identifier pour ensuite travailler de façon plus précise sur un ensemble de lectures restreint. Leur identification est cruciale. Bien entendu il faut sélectionner l'ensemble de toutes les lectures, avoir une bonne sensibilité, pour ne pas avoir de faux négatif. Mais aussi il faut avoir une bonne précision, limiter le nombre de faux positifs qui pourraient influer négativement sur les corrections mises en œuvre dans la troisième étape. Pour identifier les lectures d'intérêt, notre méthode propose d'utiliser deux outils différents : BLAST [Alt+90] et minimap2 [Li18]. Ils ont été choisis pour leur approche complémentaire dépendamment de la séquence de référence fournie. Une étude du résultat de la sélection des lectures avec ces deux outils est présentée section 4.2. Il est à noter qu'une certaine permissivité est tolérée dans le choix des lectures à cette étape. En effet, d'autres critères de sélection seront appliqués sur les lectures par la suite pour éliminer les faux positifs.

Sélection avec pour séquence de référence le gène seul. Le programme BLAST est utilisé dans sa version megablast. Sa spécificité est d'utiliser des graines d'alignements plus longue que blastn, respectivement 28 contre 11. Ce qui le rend plus adapté à la recherche de séquence très similaire. megablast aligne la séquence de référence du gène sur la base de données des lectures et après indexation de celle-ci. La construction ne se fait que pour un gène. Il a été sélectionné pour sa bonne sensibilité.

Sélection avec pour séquence de référence le génome entier. Avec un génome complet, la recherche avec megablast serait trop consommatrice en temps. Nous proposons ici d'utiliser minimap2 avec son option d'alignement 'splicé' pour aligner des lectures sur la séquence du génome de référence de l'espèce. L'alignement nécessite une seule indexation du génome pour aligner l'ensemble des séquences. On sélectionne ensuite les lectures qui ont été alignées contre le locus du gène d'intérêt. L'utilisation de minimap2 a été motivée par sa capacité à traiter les gènes paralogues.

Filtrage des lectures. Quelque soit l'outil, les lectures sélectionnées ne sont conservées que si le pourcentage de couverture de l'alignement sur la lecture est supérieur à un seuil donné (50% par défaut). La couverture est calculée à partir des hits avec corrections des chevauchements. On considère en effet que si moins de la moitié de la lecture est alignée sur la référence alors elle ne possède pas une similarité satisfaisante avec la séquence du gène d'intérêt.

3.2.2 Correction des lectures

Les séquences des lectures sélectionnées à l'étape précédente sont corrigées grâce à l'outil de correction de séquences des longues lectures isONcorrect dcrit section 2.3.2. Cette correction permet d'améliorer la précision des étapes suivantes du pipeline en diminuant les erreurs de séquençage. La question de savoir s'il faut faire une correction ou non est difficile à trancher. En effet, corriger c'est potentiellement éliminer des petits signaux, typiquement ceux de l'épissage alternatif en $3'$ ou $5'$. C'est pourquoi cette étape reste optionnelle. Une discussion sur ce point est réalisée dans le chapitre 4.

3.2.3 Pré-filtrage des structures prédites.

Avant de commencer les étapes de raffinement des résultats pour la prédiction des isoformes, on utilise une première fois exonerate pour réaliser un filtrage des données brutes de son alignement. L'objectif est d'enlever des lectures faussement séléctionnées avant de passer aux étapes suivantes. En effet, la rétention d'erreur dans le set de lectures peut affecter la qualité des prédictions finales, notamment au cours de l'étape de correction des lectures.

Choix de filtrage des lectures mono-exoniques

Dans le pipeline d'analyse de RNA-tailor, nous choisissons d'exclure les lectures dont l'alignement splicé donne lieu à un alignement sans intron, lectures qu'on qualifie de lecture mono-exonique. Notre intérêt se porte exclusivement sur les gènes qui présentent des dynamiques d'épissage, impliquant que ces gènes et leurs transcrits exprimés comprennent plusieurs exons. Nous estimons donc que ces lectures n'apportent aucune information supplémentaire sur la dynamique d'épissage du gène. De plus, dans le contexte du séquençage de troisième génération, ces lectures sont souvent des contaminations ou résultent d'artefacts de séquençage incomplet, offrant ainsi une information partielle. Cependant, certains gènes peuvent présenter un événement d'épissage aboutissant à un transcrit alternatif mono-exonique. Pour ces cas, il est possible de désactiver cette option dans le pipeline afin de traiter également ces situations.

Choix de filtrage des structures introniques

Une seconde étape de filtrage est réalisée pour éliminer les lectures dont la structure intronique est trop différente de celle des autres lectures. Chaque intron est évaluée en fonction de son nombre d'occurrences dans l'ensemble de données. Si une structure intronique est unique ou présente moins de 2% du total des lectures, les lectures la contenant sont supprimées.

L'analyse à l'échelle des introns permet d'identifier les mauvais appariements des jonctions entre elles, tout en conservant les jonctions exoniques correctes si elles sont supportées par d'autres appariements introniques. Cette approche permet de distinguer le bruit du signal au niveau structurel. Cette méthode a été motivée par l'analyse de la distribution des introns prédits en sortie de exonerate, comme démontré en figure 4.9.

3.3 Étape de raffinement des alignements

En entrée de cette partie du pipeline, on dispose d'un ensemble de lectures homogène, dont les séquences ont été auto-corrigées par isONcorrect, puis alignés par exonerate version est2genome.

L'objectif de cette partie de notre méthode est d'affiner la prédiction des bornes génomiques des jonctions intron-exon à partir des isoformes prédits en sortie de exonerate.

Dans un premier temps, on étudie à nouveau l'homogénéité structurelle de l'ensemble de lectures sélectionnées. On fait l'hypothèse que cet ensemble provient de l'expression d'un même gène est donc partage une même structure interne. On cherche donc à identifier, puis supprimer, les lectures potentiellement inconsistantes avec la structure globale de notre jeu de données. Ensuite, conscient des limites de l'algorithme de exonerate, on applique une méthode de détection des erreurs d'alignement des lectures au niveau des jonctions intron-exon. Pour corriger ces erreurs on utilisera exonerate en version alignement local. Enfin, après une seconde vérification de l'homogénéité structurelle globale de notre jeu de données, la méthode de lissage des bordures va permettre d'obtenir des bornes génomiques de confiance pour les jonctions exoniques. La méthode est destinée à pallier aux dernières erreurs d'alignement dues aux erreurs de séquençage.

3.3.1 Définitions préliminaires

On notera ℓ la longueur de la séquence de référence g (du gène d'intérêt) et m le nombre de lectures r_i, $1 \leq i \leq m$, du jeu de données (m est utilisé de manière générique pour désigner le nombre de lectures du jeu de données à une étape donnée du pipeline).

Les alignements splicés fournissent pour chaque lecture des sous-alignements correspondant aux parties identifiées comme exoniques par exonerate. On construit une *matrice binaire* qui a pour objectif de répertorier tous ces sous-alignements pour l'ensemble des lectures. On travaillera ensuite sur cette matrice pour identifier et corriger les jonctions intron-exon.

Définition 1 *La* matrice binaire *est une matrice M de 0/1 de taille $\ell \times m$ telle que, pour chaque lecture r_i, et chaque position p,*

$$M(p,i) = \begin{cases} 1 & \textit{si la position } p \textit{ de } g \textit{ est alignée sur } r_i \\ 0 & \textit{sinon} \end{cases}$$

Un exemple de matrice binaire est donné Figure 3.2.

On introduit maintenant la notion de *point de jonction*, qu'on déclinera au niveau des lectures et du gène. Les points de jonction correspondent à un passage d'un intron à un exon ou d'un exon à un intron dans les alignements splicés.

Définition 2 *Un* point de jonction entrant sur une lecture *est un couple de positions (p,i), $1 \leq p \leq \ell$, $1 \leq i \leq m$, tel que $M(p,i) = 1$ et $M(p-1,i) = 0$.*

Définition 3 *Un* point de jonction sortant sur une lecture *est un couple de positions (p,i), $1 \leq p \leq \ell$, $1 \leq i \leq m$, tel que $M(p,i) = 1$ et $M(p+1,i) = 0$.*

Définition 4 *Un* point de jonction entrant sur le gène e_p *correspond à une position p, $1 \leq p \leq \ell$, telle qu'il existe au moins un i, $1 \leq i \leq m$, tel que (p,i) est un point de jonction entrant.*

Définition 5 *Un* point de jonction sortant sur le gène s_p *correspond à une position p, $1 \leq p \leq \ell$, telle qu'il existe au moins un i, $1 \leq i \leq m$, tel que (p,i) est un point de jonction sortant.*

Les points de jonctions sont caractérisés par leur position mais aussi par le nombre de points de jonctions sur les lectures qui le supporte.

Définition 6 *Soit j_p un point de jonction sur le gène à la position, sa* couverture*, notée couverture(j_p) le nombre de lectures i tel que $M(p,i) = 1$ et $M(p-1,i) = 0$ si j_p est entrant, ou tel que$M(p,i) = 1$ et $M(p+1,i) = 0$ si j_p est sortant.*

La figure 3.2 illustre ces notions sur un exemple simple. On notera qu'à une position sur le gène, il peut exister à la fois un point de jonction entrant et sortant (voir Figure 3.4 pur un exemple).

On qualifie maintenant les points de jonction de manière à identifier ceux qui sont soutenus par plusieurs lectures de ceux qui ne le sont pas. Cette notion sera utilisée pour la partie filtrage de l'étape 3 du pipeline.

Définition 7 *Un point de jonction (entrant ou sortant) j_p sur le gène est* solide *si couverture$(j_p) \geq 2$.*

Définition 8 *Un point de jonction (entrant ou sortant) sur le gène est* fragile *s'il n'est pas solide.*

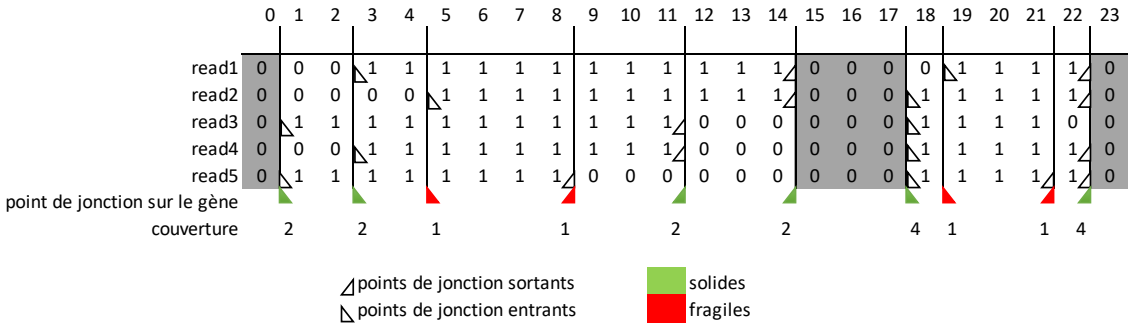


FIGURE 3.2 – Illustration du vocabulaire défini autour de la matrice binaire pour 5 lectures. Les nombres du haut indiquent la position sur la séquence de référence. Il y a un '1' si la position correspondante de la référence a été alignée sur la lecture. Les points de jonction d'entrée et de sortie sont indiqués dans les lignes correspondantes par des triangles (leur nombre est indiqué pour chaque point de jonction). Les points de jonction solides sont colorés en vert, tandis que les points de jonction fragiles sont colorés en rouge.

Notons que le seuil de support fixé à 2 ici peut être modifié. La figure 3.2 illustre ces notions.

Après le calcul des points de jonctions, la matrice binaire permet de réaliser la détermination des *blocs*. On définit un bloc comme un segment d'alignement sur la référence qui est homogène en composition de lectures. Le début et la fin d'un bloc sont marqués par l'ajout ou la perte d'une lecture pour ce segment. Le découpage en blocs permet de mettre en évidence des zones d'alignement sur la référence qui sont partagées par un sous-ensemble de lectures.

Définition 9 *Soit $j_1, \ldots, j_p$ la liste ordonnée de 5' vers 3' des points de jonction du gène g (s'il existe un point de jonction entrant et sortant à la même position, le point de jonction entrant est ordonné avant le point de jonction sortant). Il existe un* bloc *pour tout couple de points de jonctions successifs (j_{k-1}, j_k), $2 \leq k \leq p$. Il existe également, le cas échéant, un bloc entre le début du gène et j_1, et entre j_p et la fin du gène.*

La figure 3.3 illustre comment on passe des alignements splicés aux blocs pour un ensemble de lectures. Ainsi un bloc est caractérisé par les deux points de jonctions à ses extrémités. Ces points de jonction induisent les positions du bloc sur le gène. Le couple de positions de *début* et *fin* sur le gène est déterminé ainsi :

		nature de j_{k-1}	
		entrant	sortant
nature de j_k	entrant	$[j_{k-1}, j_k[$	$]j_{k-1}, j_k[$
	sortant	$[j_{k-1}, j_k]$	$]j_{k-1}, j_k]$

La figure 3.4 illustre ces cas. Notons que les blocs dont la position de début est strictement supérieur à la fin ne sont pas considérés.

Définition 10 *Soit b un bloc, de position de début x et de position de fin y, et r_i une lecture, on dira que le bloc b est* vide *pour r_i si $M(p, i) = 0$ pour tout $x \leq i \leq y$.*

On notera qu'un bloc est nécessairement vide ou « plein » pour une lecture. Sinon c'est qu'il existe un point de jonction, et donc deux blocs.

Définition 11 *Soit b un bloc, sa* couverture*, notée couverture(b) est le nombre de lectures possédant un alignement dans ce bloc.*

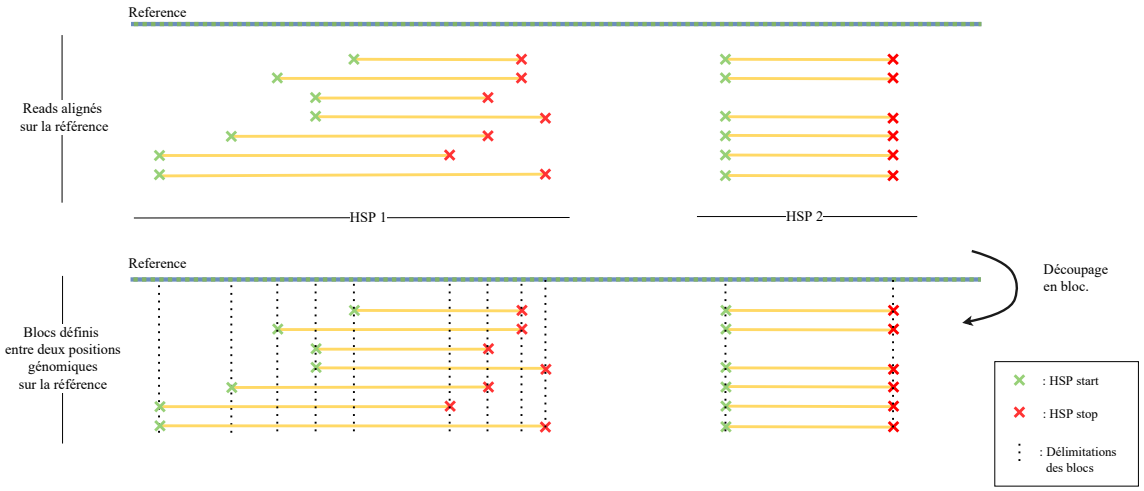
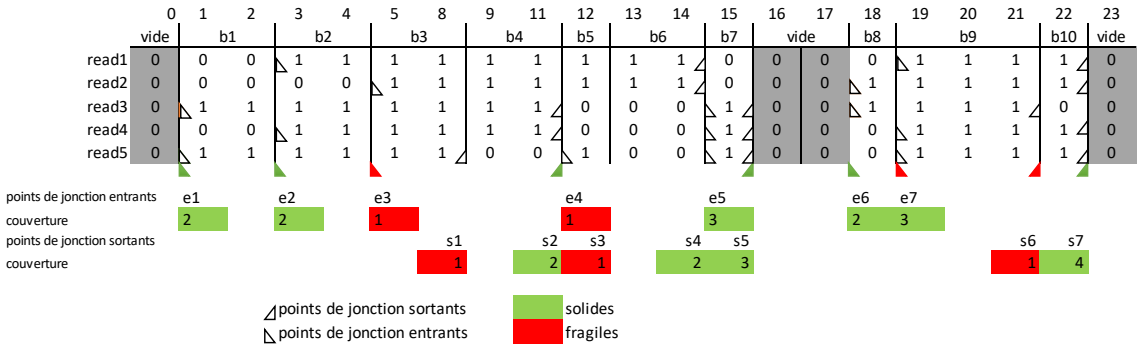


FIGURE 3.3 – Schéma explicatif de la construction des blocs à partir des données d'alignement. Chaque segment encadré de croix représente des HSPs de l'alignement de la lecture sur la référence. Cela signifie que dans la matrice binaire, on trouvera un 1 à ces positions pour cette lecture. Chaque position de début ou de fin d'un HSP définit une position de début ou de fin d'un bloc. À gauche, les HSPs révèlent une structure exonique complexe avec des débuts / fins alternatives d'exons. À droite, les HSPs révèlent une structure exonique simple (le bloc correspondant à un exon), très conservée, présent dans tous les lectures sauf le troisième.



position de j_{k-1}	j_{k-1},j_k	bloc	début	fin
1	e_1,e_2	b_1	1	2
3	e_2,e_3	b_2	3	4
5	e_3,s_1	b_3	5	8
8	s_1,s_2	b_4	9	11
11	s_2,e_4	non considéré	12	11
12	e_4,s_3	b_5	12	12
12	s_3,s_4	b_6	13	12
14	s_4,e_5	non considéré	15	14
15	e_5,s_5	b_7	15	15
15	s_5,e_6	vide	16	17
18	e_6,e_7	b_8	18	18
19	e_7,s_6	b_9	19	21
21	s_6,s_7	b_{10}	22	22
22	s_7			

FIGURE 3.4 – Illustration de la détermination des débuts et fins de bloc en fonction des points de jonction qui les caractérisent.

Remarquons que la couverture peut être calculée à partir des couvertures du bloc précédent (ou suivant) et des points de jonctions des deux blocs. La couverture permet de distinguer deux types de bloc : les blocs non-vides, avec une couverture non nulle, et les blocs vides avec une couverture nulle. Les blocs vides correspondent naturellement aux introns. Idéalement, les blocs couverts formeraient des exons. Intuitivement, plus un bloc a une couverture élevée, et plus on peut se fier à son appartenance à l'exon d'un gène (Figure 3.3). Par construction, deux blocs vides sont forcément séparés par au moins un bloc non-vide. Par contre, on peut avoir plusieurs blocs non-vides successifs.

Nous allons maintenant qualifier les blocs. Les blocs *solides* sont ceux qu'on ne remettra pas en cause, ils forment le cœur des exons tandis que les blocs *fragiles* sont ceux qui vont servir à déterminer les parties à réaligner. Soient i et j des numéros de blocs non vides, $1 \leq i \leq j \leq m$, tels que les deux propriétés suivantes sont satisfaites : le début du bloc b_i correspond à un point de jonction entrant solide et la fin du bloc b_j correspond à un point de jonction sortant solide. Alors b_i et tous les blocs suivants jusqu'à possiblement b_{j-1} de couverture supérieure ou égale à couverture(b_i) sont solides, et b_j et tous les blocs précédents b_j jusqu'à possiblement b_{i+1} de couverture supérieure ou égale à couverture(b_j) sont solides. Les autres blocs non vides sont fragiles. La détermination des blocs solides est illustrée par la figure 3.5. Notons que les suites de blocs en 5' et 3' (i.e. avant le premier

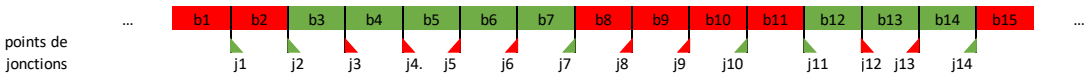


FIGURE 3.5 – Schéma explicatif de la détermination des blocs solides. Les points de jonction entrants (resp. sortants) sont représentés par des triangles orientés à gauche (resp. orientés à droite). Les verts sont solides, les rouges fragiles. Les suites de blocs solides sont surlignés en vert, les blocs fragiles en rouge. Pour la première suite, on a cherché le premier bloc se terminant sur un point de jonction sortant solide (bloc b_7 avec j_6) puis a remonté les blocs jusqu'à trouver le premier bloc avec un point de jonction entrant solide (bloc b_3 avec j_2). On n'illustre pas ici la détermination des blocs solides au sein de chaque suite mais on pourrait imaginer que b_4, b_5, b_6 soient fragiles si leur couverture était trop différente de celles de b_3 et b_7.

bloc vide, et après le dernier bloc vide) sont des cas particuliers qui ne suivent pas cette règle. En 5', la suite de blocs solides commence au premier bloc et se termine au premier bloc ayant point de jonction sortant solide. En 3', c'est l'inverse, la suite de blocs solides commence au bloc ayant le dernier point de jonction entrant solide et se termine au dernier bloc. L'algorithme 1 illustre la détermination des blocs solides.

La qualification des blocs par un statut permet d'identifier s'ils sont éligibles pour le réalignement. L'identification des zones de réalignements est unique à chaque lecture.

Définition 12 *Une* zone de réalignement *(un groupe de blocs réalignables) pour une lecture r est une une suite de blocs telle que :*

- *la suite est encadrée par deux blocs solides, tous les deux non vides pour r ;*
- *au moins un bloc est vide pour r ;*
- *au moins un bloc n'est pas vide pour r ;*
- *tous les blocs non vides pour r sont fragiles.*

Pour chaque zone de réalignement on déduit la séquence de la lecture correspondante comme la concaténation des séquences de la lecture pour chaque bloc, et on appellera *longueur de la zone de réalignement pour la lecture* la longueur de cette sous-chaîne de la lecture. Cette séquence est celle susceptible d'être mal alignée. La région du gène contre laquelle le réalignement peut s'effectuer est bornée par deux ancres formées par les deux blocs solides précédant et suivant la zone de réalignement. La figure 3.6 illustre différents cas de figure.

Algorithme 1 : Détermination des blocs solides.

Entrée : La liste ordonnée des blocs
`// Calcul pour les blocs en 5' (avant le premier bloc vide)`
Soit B_d le dernier bloc avant le premier bloc vide du premier exon et ayant un point de jonction sortant solide.
$k = d - 1$
while $k > 0$ *et* B_k *possède un point de jonction sortant fragile et est contiguë à* B_{k+1} **do**
  rendre B_k solide ; $k - -$
end
`// Calcul pour les blocs en 3' (après le dernier bloc vide)`
Soit B_p le premier bloc après le dernier bloc vide du dernier exon ayant un point de jonction entrant solide.
$k = p + 1$
while $k <= m$ *et* B_k *possède un point de jonction entrant fragile et est contiguë à* B_{k-1} **do**
  rendre B_k solide ; $k + +$
end
`// Calcul pour les blocs entre le premier bloc vide et le dernier bloc vide`
Soit t l'indice du premier bloc après le premier bloc vide
while $t \leq$ *indice du dernier bloc* **do**
  Soit j le plus petit indice supérieur ou égal à t tel que $s_j > 1$
  Soit i le plus grand indice inférieur ou égal à j et supérieur ou égal à t tel que $e_i > 1$
  `// ce qui assure que pour tout` k, $i < k < j$, $e_k < 2$ `et` $s_k < 2$)
  `// ensuite, pour savoir quels sont les blocs solides entre` i `et` j
  $k_1 = i$; **while** $k_1 < j$ *et* $cover(B_{k_1}) >= cover(B_i)$ **do**
    rendre B_{k_1} solide ; $k_1 + +$
  end
  $k_2 = j$ **while** $k_2 >= k$ *et* $cover(B_{k_2}) >= cover(B_j)$ **do**
    rendre B_{k_2} solide ; $k_2 - -$
  end
  $t = j + 1$
end
Recommencer avec le prochain point de jonction sortant solide.

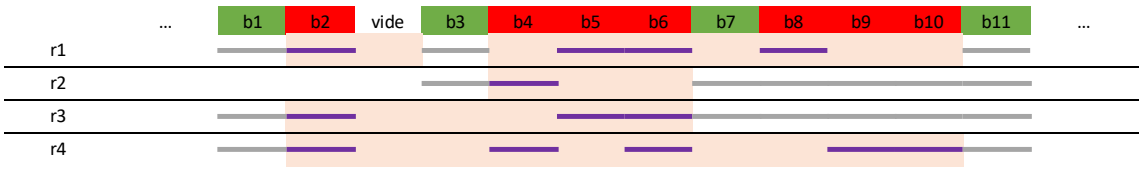


FIGURE 3.6 – Illustration schématique de la détermination des zones de réalignement. Pour chacune des 4 lectures, figurent en trait plein les parties de la lecture alignée dans chacun des blocs, en violet les segments correspondant aux zones de réalignement. Surlignés en saumon les segments génomiques contre lesquels chaque zone de réalignement sera réalignée. Les blocs verts sont solides tandis que les blocs rouge sont fragiles. Par exemple, la séquence du bloc $b2$ de la lecture $r1$ sera réaligné contre la portion génomique contenue entre $b1$ et $b3$, partie intronique comprise. De même, pour $r4$ les séquences des blocs $b2$, $b4$, $b6$, $b9$ et $b10$ sont réalignables sur la même zone. Les séquences génomiques de chaque bloc seront concaténées en une séquence qui sera réalignée sur la portion du gène comprise entre $b1$ et $b11$.

3.3.2 Pré-filtrage et filtrage des points de jonctions.

Le filtrage des lectures vise à limiter la présence de lectures qui ne proviennent pas du gène d'intérêt, que l'on considère comme des faux positifs de l'étape de sélection. On suppose que ces lectures possèdent une structure interne différente de celle du gène d'intérêt, et donc de celle observée dans les autres lectures. Pour les identifier, on étudie l'homogénéité des positions des points de jonctions entre les lectures.

Deux étapes de filtrage interviennent dans le pipeline. La première, qu'on appelle pré-filtrage, intervient avant la correction des alignements et est conditionnelle. La seconde, qu'on appelle filtrage, intervient après.

Afin d'identifier les lectures à enlever, on introduit la notion de point de jonction isolé et de lecture suspecte.

Définition 13 *Un point de jonction* (p, i) *d'une lecture est* isolé *si la distance maximale avec les autres lectures est supérieure à 20 nucléotides.*

Définition 14 *Une lecture est* suspecte *si plus de 25% de ses points de jonctions sont isolés.*

Le filtrage se fait en deux étapes : on identifie d'abord les lectures suspectes, puis on décide ou non de les enlever du jeu de données suivant l'étage de filtrage.

Pré-filtrage. Le pré-filtrage permet une élimination précoce des lectures conditionnées à une évaluation de la possibilité de replacer correctement des points de jonction suite à un réalignement. L'intérêt de cette élimination conditionnelle est de conserver les lectures de vrais positifs ayant connu un problème d'alignement par exonerate. Le pré-filtrage parie sur la capacité de l'étape de correction des alignements à racheter ces lectures de vrais positifs. Dans le cadre du pré-filtrage, une lecture suspecte est conservée si elle possède une ou plusieurs zones de réalignements qui permettrait de convertir des points de jonctions isolés en suffisamment de points de jonctions solides pour franchir le critère de 25% de points de jonctions solides. Pour les lectures suspectes les zones de réalignements possibles sont d'abord calculées. Puis pour chaque zone de chaque lecture r, définissant une sous-chaîne du s de r, on évalue s'il existe une possibilité de réaligner cette zone sur une partie du gène déjà couverte par une autre lecture. On prend comme critère que la partie couverte par l'autre lecture doit être de longueur supérieure à la longueur de s. Si c'est le cas, alors les points de jonction isolés sont considérés comme non isolés, et la qualité de la lecture est ré-évaluée. Si elle est non suspecte, alors la lecture est conservée. La figure 3.7 illustre le rachat de lecture lors du processus de pré-filtrage.

Filtrage. Intervenant après le pré-filtrage, il permet d'éliminer les lectures qui n'ont pas pu être rachetées à l'étape précédente. Toutes les lectures suspectes sont supprimées.

3.3.3 Identification et correction des erreurs d'alignements

Le réalignement de certaines parties des lectures permet de corriger des erreurs d'alignements sur les bordures d'un ensemble de blocs. La Figure 3.8 en donne une illustration. Les longueurs des zones de réalignement sont variables, allant de quelques bases à plusieurs dizaines. Notons immédiatement que le réalignement des zones de réalignement en début et fin de lecture n'est pas opéré. Ces alignements étant souvent réalisés sur des portions génomiques de grande taille (tout le début du gène, ou toute la fin du gène), le coût de calcul d'alignement est très important, pour un apport en information négligeable.

Nous adoptons deux stratégies différentes suivant la longueur des zones de réalignement. Au dessus de 25 nucléotides, la sous-chaîne de la lecture est réalignée par exonerate dans sa version d'alignement local sur la sous-chaîne du gène correspondante. Cette méthode est appelée *réalignement doux* en opposition au *réalignement dur* pour les séquences plus

FIGURE 3.7 – Illustration du pré-filtrage. Sur ce schéma, les portions grises des lectures correspondent aux blocs solides. Ces portions sont supportées par des jonctions solides identifiées par un carré vert. Les portions jaunes correspondent aux blocs fragiles. Ils sont soutenus par des jonctions fragiles identifiées par un carré rouge. Les zones violettes correspondent aux zones de réalignement identifiées pour les portions jaunes qu'elles englobent. Les lectures r1 et r2 possèdent des taux de jonctions isolées de 50% ce qui est supérieur au seuil de 25%. Ces lectures sont donc candidates pour le pré-filtrage. Elles possèdent des zones de réalignement, mais seule la zone de réalignement de r1, comprise entre $b1$ et $b10$ comprend des blocs pouvant recevoir le réalignement de $b4$ et possédant déjà une lecture comme $b5$, $b6$, $b7$ et $b9$. La lecture r1 va donc être conservée pour l'étape de réalignement. A l'inverse, la lecture r2 ne possède pas de bloc dans sa zone de réalignement pouvant recevoir la concaténation de $b7$ et $b9$ et possédant une lecture déjà alignée. Il n'y a donc pas de possibilité de gains de point de jonctions pour r2 qui sera supprimé avant le réalignement. Pour finir, la lecture r3 possédant moins de 25% de points de jonction isolés, elle passe le filtre et sera candidate pour le réalignement.

courtes. En effet, en deçà de 25 nuléotides, on estime que la séquence nucléotidique peut s'aligner sur trop de zones différentes dans le gène. Son réalignement n'apporterait pas plus de fiabilité. Dans ce cas, on repositionne les fragments de lectures à réaligner sur la portion candidate la plus prometteuse. Néanmoins, les zones de réalignement de longueur inférieure au seuil de lissage des bordures (voir section 3.3.4, typiquement 3) ne sont pas réalignées.

Le réalignement doux est réalisé par exonerate avec son mode `affine:local` et nous avons fait le choix de ne retourner que le meilleur HSP. Si la position du nouvel alignement est similaire à l'ancienne, on invalide le réalignement. Une position est similaire si elle est inclue dans l'ancien alignement dans une fenêtre de deux nucléotides. Dans ce cas, le premier alignement proposé par exonerate était déjà le plus vraisemblable. Lors d'un réalignement local, il est possible que des sous-parties de la séquence se réalignent en plusieurs hits. Ces hits ne sont pas contraints de conserver le même ordre que la séquence d'origine, ni d'être alignés à des endroits différents. Les hits peuvent donc se retrouver dans le désordre et se chevaucher. Pour être accepté, le nouvel alignement doit respecter la colinéarité de la séquence d'origine et ses différents hits ne doivent pas se chevaucher.

La méthode du réalignement dur à pour objectif de repositionner des fragments du lecture trop petits pour être réalignés localement. Pour choisir la position de réalignement la plus probable, la méthode sélectionne dans un premiers temps les groupes de blocs (sous-entendus contigus) assez grands pour accueillir les fragments à réaligner. Ces groupes de blocs doivent nécessairement contenir un bloc solide. Pour une zone de réalignement de longueur p, un groupe de blocs de longueur totale n est accepté si $|n - p|$ est inférieur à $2 \times$ la valeur du seul de lissage. Pour choisir vers quel groupe de blocs réaligner, on va donner un poids à un bloc puis choisir la combinaison de blocs maximisant la somme des poids de ses blocs. Cela permet de rechercher la suite de blocs d'accueil étant à la fois assez peuplés en lectures et ayant une dynamique d'alignement correspondant à la taille du fragment à réaligner.

Définition 15 *Le* poids *d'un bloc b, possédant e points de jonctions entrants et s points de jonctions sortant est défini par :*

$$w(b) = e + s$$

Définition 16 *Pour une séquence de blocs contiguës $b_i, \ldots, b_j$, le* score d'attractivité *est*

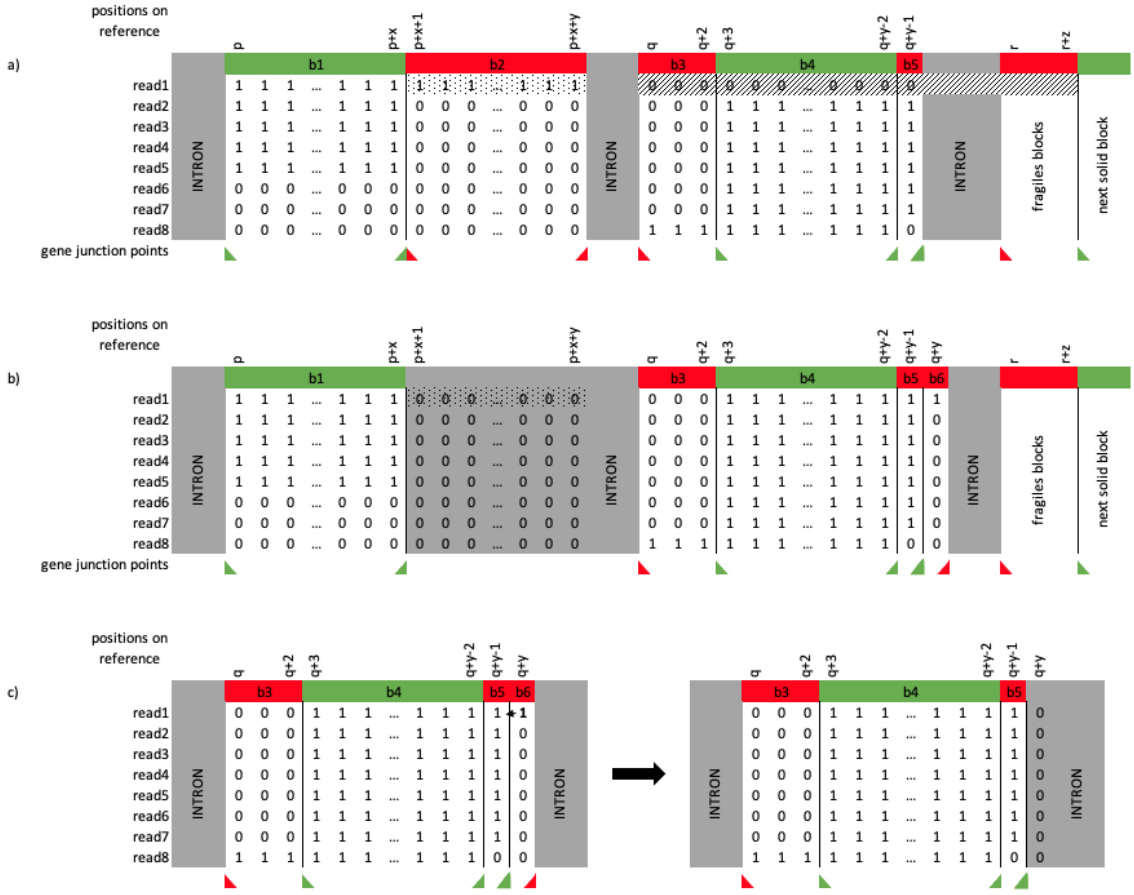


FIGURE 3.8 – Illustration du réalignement et du lissage des bordures. Vert/Rouge : blocs solides/fragiles. Triangle vert/Rouge : points de jonction entrant/sortant fragiles/solides. 1 et 0 montrent la matrice binaire. a) le bloc solide gauche b_1 a une longueur de x et le bloc fragile suivant a une longueur de y. Seule la première lecture est alignée contre la référence dans ce bloc fragile. Le bloc solide droit b_4 a une longueur de $y-2$. Comme il y a de la place pour réaligner presque toute la sous-séquence de la première lecture dans le bloc solide b_4, cela est fait. En fonction de la valeur de y, l'étape de réalignement dur ou doux est appliquée. En cas de réalignement dur, la zone de réalignement doit être comprise entre les positions $p+x+1$ et $r+z$. b) après ce réalignement, le bloc fragile entre les positions $p+x+1$ et $p+x+y$ est vide et sera fusionné avec l'intron. La matrice binaire a maintenant des 1 dans le bloc solide droit b_4 de longueur $y-2$. Comme le segment réaligné avait une longueur de y, un nouveau bloc fragile b_6 de longueur 1 a été créé. c) focus sur les blocs b_3, b_4, b_5, b_6. Un lissage des bordure est appliqué pour corriger la fin de l'alignement de la lecture 1 une position avant. En effet, le lissage des bordures s'appliquant à partir de b_4 en aval, ne s'applique pas à b_5 car il est supporté par une jonction solide. Il s'applique ensuite à partir de b_5 sur b_6 qui possède une jonction sortante fragile. La position de fin d'alignement de la lecture 1 dans b_6 est donc supprimée car réassignée en b_5.

défini par :

$$score(i,j) = \sum_{k=i}^{j} w(b_k)$$

3.3.4 Lissage des bordures

Alors que le réalignement corrige des suites de blocs de longueur cumulée supérieure à 3 nucléotides, le lissage des bordures est la méthode de correction appliquée lorsque la longueur de cette suite est de 2 nucléotides ou moins. Corriger des portions de taille inférieure à celle d'un codon a pour objectif de diminuer les problèmes de « *frameshifts* » et d'augmenter la qualité des prédictions de RNA-tailor aux sites d'épissage. La méthode tente donc de distinguer les blocs, représentant une information, des autres. Elle doit réaliser le compromis de réduire la variabilité présente dans le jeux de donnée tout en préservant les vrais événements d'épissage. Pour cela, elle s'appuie sur la liste des blocs possédant un point de jonction solide (entrant ou sortant) en amont et en aval de chaque début et fin de bloc solide. En amont, le lissage s'applique au niveau de tous les blocs possédant des jonctions solides (B_{js}) rencontrées jusqu'au prochain bloc vide ou sortant solide. Pour chacun de ces blocs B_{js}, la méthode cherche s'il existe un ou deux blocs à jonction entrante fragile (B_{jf}) dans une fenêtre de 2 nucléotides en amont. S'il existe de tels blocs B_{jf}, leurs alignements sont considérés comme peu vraisemblables et sont corrigés vers la position du bloc B_{js}. On cherche à décider si chacun de ces blocs représente ou non une information de jonction d'épissage à conserver. Pour cela, on choisit de garder uniquement les blocs possédant un point de jonction solide, les autres seront lissés. Ils intégreront le bloc adjacent b_i. La partie c) de la figure 3.8 donne une illustration de l'objectif et de la méthode de poursuite du lissage.

3.4 Détermination des isoformes

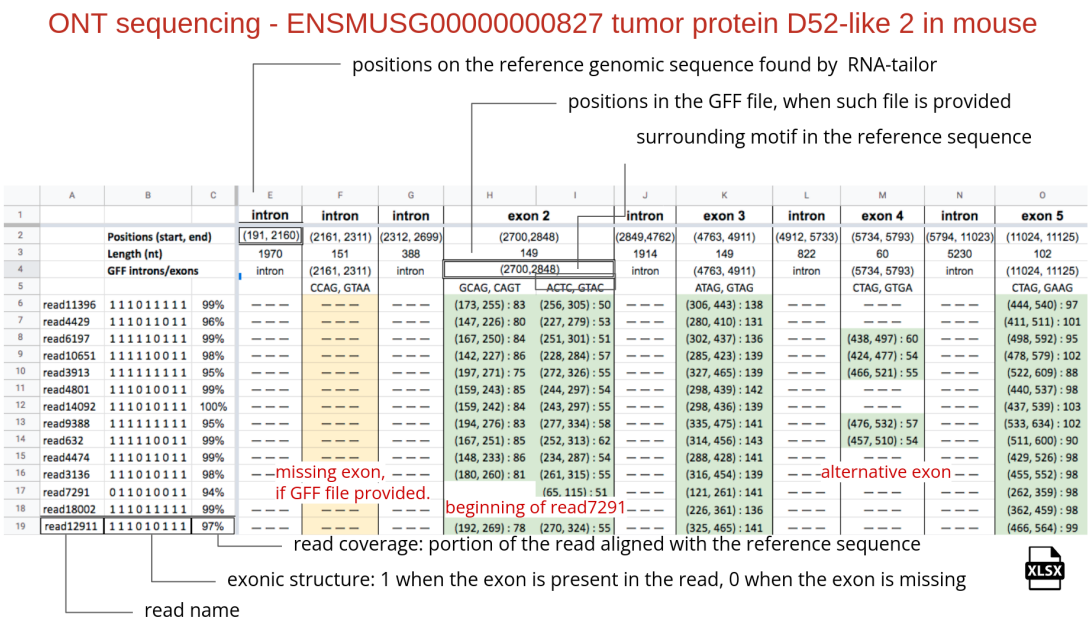
L'objectif de cette étape est de proposer un modèle de structure exonique des isoformes à partir de la structure en bloc des lectures. Pour cela, on détermine pour chaque lecture les bornes de ses exons et introns à partir des blocs qui le composent. Les bornes introniques de chaque lecture sont alors comparées. Si deux lectures possèdent le même nombre d'introns et que ceux-ci ont les mêmes bornes génomiques alors ces deux lectures supportent un même isoforme. Etudier la structure intronique permet de contourner le problème de complétude des longues lectures (voir section 2.2.2). Les TTS et TSS des isoformes composés de plusieurs lectures sont égales à leur valeur extrême parmi les lectures.

3.5 Implantation et sorties de RNA-tailor

Après la détermination des isoformes, quatres sorties standardes sont effectuées par RNA-tailor : une feuille de calcul (au format xlsx), un fichier JSON qui résume notre structure de données (la composition en blocs de toutes les lectures) et qui permet également un post-traitement des lectures (voir section suivante), et deux fichiers au format GTF qui publie l'état des prédictions d'isoformes après l'exécution de RNA-tailor.

Un exemple annoté de l'export de notre structure de donnée sous la forme de tableur est visible sur la figure 3.9. Ce tableur résume notre structure de données et peut-être exporté à partir de celle-ci entre chaque étape du pipeline. Elle permet ainsi de visualiser les effets de chaque étape du pipeline à partir des alignements initiaux des lectures par exonerate. Cette représentation permet de comprendre l'alignement des lectures et les mécanismes d'épissage identifiés.

La prédiction de certains isoformes n'est supportée que par un très faible nombre de lectures. Cela permet de filtrer les isoformes en deux catégories selon la fiabilité de leur prédiction : les isoformes dits *solides* et dits *fragiles*.



				intron	intron	intron	exon 2		intron	exon 3	intron	exon 4	intron	exon 5
	Positions (start, end)			(191, 2160)	(2161, 2311)	(2312, 2699)	(2700,2848)		(2849,4762)	(4763, 4911)	(4912, 5733)	(5734, 5793)	(5794, 11023)	(11024, 11125)
	Length (nt)			1970	151	388	149		1914	149	822	60	5230	102
	GFF introns/exons			intron	(2161, 2311)	intron	(2700,2848)		intron	(4763, 4911)	intron	(5734, 5793)	intron	(11024, 11125)
					CCAG, GTAA		GCAG, CAGT	ACTC, GTAC		ATAG, GTAG		CTAG, GTGA		CTAG, GAAG
read11396	1 1 1 0 1 1 1 1 1	99%		– – –	– – –	– – –	(173, 255) : 83	(256, 305) : 50	– – –	(306, 443) : 138	– – –	– – –	– – –	(444, 540) : 97
read4429	1 1 1 0 1 1 0 1 1	96%		– – –	– – –	– – –	(147, 226) : 80	(227, 279) : 53	– – –	(280, 410) : 131	– – –	– – –	– – –	(411, 511) : 101
read6197	1 1 1 1 1 0 1 1 1	99%		– – –	– – –	– – –	(167, 250) : 84	(251, 301) : 51	– – –	(302, 437) : 136	– – –	(438, 497) : 60	– – –	(498, 592) : 95
read10651	1 1 1 1 1 0 0 1 1	98%		– – –	– – –	– – –	(142, 227) : 86	(228, 284) : 57	– – –	(285, 423) : 139	– – –	(424, 477) : 54	– – –	(478, 579) : 102
read3913	1 1 1 1 1 1 1 1 1	95%		– – –	– – –	– – –	(197, 271) : 75	(272, 326) : 55	– – –	(327, 465) : 139	– – –	(466, 521) : 55	– – –	(522, 609) : 88
read4801	1 1 1 0 1 0 0 1 1	99%		– – –	– – –	– – –	(159, 243) : 85	(244, 297) : 54	– – –	(298, 439) : 142	– – –	– – –	– – –	(440, 537) : 98
read14092	1 1 1 0 1 0 1 1 1	100%		– – –	– – –	– – –	(159, 242) : 84	(243, 297) : 55	– – –	(298, 436) : 139	– – –	– – –	– – –	(437, 539) : 103
read9388	1 1 1 1 1 1 1 1 1	95%		– – –	– – –	– – –	(194, 276) : 83	(277, 334) : 58	– – –	(335, 475) : 141	– – –	(476, 532) : 57	– – –	(533, 634) : 102
read632	1 1 1 1 1 0 0 1 1	99%		– – –	– – –	– – –	(167, 251) : 85	(252, 313) : 62	– – –	(314, 456) : 143	– – –	(457, 510) : 54	– – –	(511, 600) : 90
read4474	1 1 1 0 1 1 0 1 1	99%		– – –	– – –	– – –	(148, 233) : 86	(234, 287) : 54	– – –	(288, 428) : 141	– – –	– – –	– – –	(429, 526) : 98
read3136	1 1 1 0 1 0 1 1 1	98%		– – –	– – –	– – –	(180, 260) : 81	(261, 315) : 55	– – –	(316, 454) : 139	– – –	– – –	– – –	(455, 552) : 98
read7291	0 1 1 0 1 0 0 1 1	94%		– – –	– – –	– – –		(65, 115) : 51	– – –	(121, 261) : 141	– – –	– – –	– – –	(262, 359) : 98
read18002	1 1 1 0 1 1 1 1 1	99%		– – –	– – –	– – –			– – –	(226, 361) : 136	– – –	– – –	– – –	(362, 459) : 98
read12911	1 1 1 0 1 0 1 1 1	97%		– – –	– – –	– – –	(192, 269) : 78	(270, 324) : 55	– – –	(325, 465) : 141	– – –	– – –	– – –	(466, 564) : 99

FIGURE 3.9 – Exemple d'un export XLSX en sortie de RNA-tailor. La feuille de calcul illustre la structure d'alignement produite par RNA-tailor, chaque ligne représente une lecture identifiée par son nom, chaque colonne représente un segment d'alignement du lecture sur la séquence de référence du gène. Si une colonne est pleine, elle peut montrer une dynamique d'épissage constitutive ou un exon cassette possible si la colonne n'est pas représentée par toutes les lectures. Ici, nous avons fourni à RNA-tailor un fichier d'annotation permettant une comparaison visuelle, sur cette sortie, avec les prédictions.

Définition 17 *Un transcrit isoforme prédit par RNA-tailor est dit* solide *s'il est supporté par au moins deux lectures ou par 1% du nombre de lectures. Sinon il est dit* fragile.

3.6 Post-traitement des isoformes prédits.

En sortie de RNA-tailor, on observe qu'un grand nombre d'isoforme n'est soutenu que par une lecture mais également qu'une partie des isoformes semblent être des « match » incomplet des isoformes de référence, voir section 4.6. On cherche donc des pistes exploratoires pour affiner ces résultats et tirer davantage parti des résultats de RNA-tailor.

Ce script de post-traitement est le second script appelé lors de l'exécution de RNA-tailor. Il prend en entrée la sortie JSON de l'étape précédente. Son objectif est d'identifier les isoformes « viables » prédits à partir des alignements corrigés. C'est-à-dire donnant plus vraisemblablement lieu à la production d'une protéine fonctionnelle ou non. En effet, certaines lectures produites par le séquençage ne représentent pas des transcrits viables ou biologiquement pertinents. Parmi les transcripts d'origine se mêlent les transcrits en cours de transcription et d'épissage, les ARNs viables et les ARNs en cours de dégradation. Notre approche se base sur ses différentes caractéristiques pour identifier les transcrits les plus vraisemblables :

- correction des sites d'épissage vers les sites canoniques : on considère qu'un transcrit viable possède l'ensemble de ses sites d'épissage supportés par des sites accepteurs et donneurs canoniques ;
- **suppression des sites d'épissage faible** : on considère qu'un isoforme possédant l'unique version d'un site d'épissage, non soutenu par des sites accepteurs et donneurs canoniques comme une lecture peu fiable ;
- **classification par recherche de codon START/STOP** : on considère qu'un transcrit est viable, s'il possède une unique ORF (Open Reading Frame) de longueur

maximale supérieure à 150 nucléotides.
— **Regroupement par séquences introniques incluses** : on considère qu'un isoforme A ayant l'entièreté de sa structure intronique incluse dans celle d'un isoforme B plus grand, représente une version dégradée de B et est donc fusionné avec B.
— **Filtrage par support de lecture** : on considère que plus la structure d'un isoforme est supportée par un grand nombre de lecture et plus il est fiable.
— **Filtre sur la longueur des lectures** : on considère que les isoformes les plus courts sont ceux qui ont subit le plus de dégradation. Ainsi plus un isoforme est long et plus il a de chance de représenter la structure d'un transcript entier.

3.6.1 Correction des sites d'épissage vers les sites canoniques

La correction des sites d'épissage vers les signaux canoniques est réalisé pour chaque lecture indépendamment. Pour chacun de ses sites d'épissage, suivant s'il est donneur ou accepteur, la méthode cherche un signal d'épissage dans une fenêtre de deux nucléotides avant et après la position prédite. Si le bon signal de d'épissage est identifié dans cette fenêtre, alors on considère qu'il y a eu erreur d'alignement. La position de la jonction sera remplacée par la position adjacente au signal d'épissage.

3.6.2 Regroupement par ORF

Même si la protéine produite par les isoformes n'est pas accessible, on peut réfléchir en terme d'ORF à partir de ces derniers. Le but est d'apporter une information complémentaire qui permet de créer du lien entre structures introniques. En effet, deux isoformes possédant une structure intronique différente peuvent coder pour la même protéine. Ainsi, deux isoformes sont fusionnés s'ils possèdent la même ORF car ils n'apportent pas plus d'informations pour étudier les traductions probables en protéine. On suppose que l'ORF la plus probable pour un isoforme est celle de longueur maximale. Les ORF sont calculées en utilisant le package "orffinder" [Cho21]. Un exemple de regroupement par ORF est visible en figure 4.19a.

3.6.3 Regroupement par séquences introniques incluses.

Les lectures produites par séquençage de troisième génération sont fortement affectées par des dégradations en début et fin de lectures. Même si ces lectures sont de tailles différentes, elles peuvent provenir d'un même isoforme. On parle alors de « *match* » incomplet 4.10. Ainsi la présence des exons en début et fin d'isoformes peut être impacté par cette dégradation et a pour conséquence de créer des isoformes trop courts. Partant de ce constat, on fait l'hypothèse que si un isoforme partage l'ensemble de ses jonctions introniques avec un autre isoforme plus long, alors ils sont inclus l'un dans l'autre. Cette hypothèse est très forte et néanmoins on a souhaité tester son impact sur des données prédites et sur la base de données de référence 4.1. Pour réaliser un clustering basé sur l'inclusion de structures introniques, on va associer à chaque lecture un vecteur binaire. La taille de ce vecteur est égale au nombre d'introns différents observés dans un ensemble d'isoformes. Chaque position de ce vecteur correspond à un intron et est égale à 1 si l'isoforme possède cet intron et 0 sinon. Le critère d'inclusion se définit tel que : un vecteur b est inclus dans un vecteur a si a possède un 1 à toutes les positions où b possède un 1. Ce qui revient à écrire que $(b \wedge a) = b$. Voici un exemple de cette approche d'inclusion avec des vecteurs binaires : soit deux vecteurs binaires, a and b, chacun de longueur n. Chaque position du vecteur binaire correspond à un intron contenu dans nos prédictions, représenté par un couple de positions nucléotidiques.

Exemple de vecteurs inclus : $a = [1, 1, 0, 1]$, $b = [1, 0, 0, 1]$. Ici b est inclus dans a, car a contient 1 à toutes les positions où b contient 1.

Exemple de vecteurs non inclus : $a = [0, 1, 1, 0]$, $b = [0, 1, 0, 1]$. Inversement, b n'est pas inclus dans a, car a ne possède pas toutes les positions égale à 1 de b.

L'hypothèse derrière la méthode de l'inclusion étant très forte, on cherche à la rendre plus sensible en ajoutant des règles de non-inclusion pour protéger des structures introniques d'intérêt vis-à-vis de l'inclusion. Pour cela, on introduit l'inclusion avec sensibilité au ORF et l'inclusion avec sensibilité au statut UTR.

3.6.4 Inclusion avec sensibilité aux ORF

On observe que certains isoformes courts, codant pour des protéines plus courtes, sont inclus dans des isoformes plus longs, codant pour des protéines plus longues. Dans ces cas on observe que les structures plus courtes permettent de faire apparaître un codon start alternatif, générant une protéine différente. C'est donc une information viable que l'on souhaite conserver en ajoutant la sensibilité aux ORF. Cette méthode compare les ORF de taille maximale de l'isoforme court inclus et l'isoforme long. Si l'ORF maximale de l'isoforme court partage le même codon START ou STOP que l'isoforme, tout en ayant un codon START ou STOP alternatif sur la même phase, alors il est considéré comme un isoforme alternatif valable et est conservé. Un exemple est visible en figure 4.19a.

3.6.5 Inclusion avec sensibilité au statut UTR

La motivation derrière l'ajout de cette règle est similaire à celle de la sensibilité aux ORF 3.6.4, dans le sens où l'on cherche à reconnaître des isoformes ayant des codons START ou STOP alternatifs. La différence est qu'ici, on utilise les prédictions de RNA-tailor pour la reconnaissance d'UTR. On fait l'hypothèse qu'une lecture possédant un statut UTR en début ou fin d'alignement soutient la structure d'un isoforme alternatif à un isoforme plus long, dans lequel il serait inclus. On définit le statut UTR d'un bloc dans RNA-tailor tel que : l'alignement d'une lecture L dans un bloc fragile possède un statut UTR, si L ne possède pas d'alignement dans un bloc solide en amont du côté 5' de cet alignement ou en aval du côté 3'. Un exemple est également visible en figure 4.19a.

3.6.6 Filtrage par support de lecture.

Le post-traitement inclus le filtrage par support de lecture de la même manière que dans le script principale de RNA-tailor. Cela permet de modifier le critère de support à la valeur souhaitée sans avoir à relancer l'analyse.

3.6.7 Filtrage sur la longueur des lectures.

La dégradation dans les lectures issues des technologies de séquençage de 3$^{\text{ème}}$ génération est un phénomène connu [ŁTD16]. Elle a lieu en 5' et 3' de l'ARN. On peut donc être tenté de filtrer les isoformes soutenus par ces lectures que l'on juge trop courtes car ayant subit le plus de dégradation. A l'inverse, plus un isoforme est long et plus il a de chance de représenter la structure d'un transcript entier. Cet approche est différente de celle du filtrage sur la couverture faite dans l'étape 3.2.1. Pour mettre en œuvre un tel filtrage on évalue la distribution des longueurs des isoformes pour un gène et applique un seuil. Par défaut, la fonction filtre les isoforme qui sont parmi les 25% les plus courtes. Travailler sur la distribution de longueur permet de s'adapter au contexte de la sélection des lectures indépendamment pour chaque gène.

Conclusion

Nous avons présenté RNA-tailor, un pipeline d'analyse conçu pour traiter les longues lectures afin d'identifier les isoformes alternatifs d'épissage à l'échelle du gène. Il se divise en plusieurs étapes clés : la sélection des lectures, le filtrage des lectures contaminantes, la correction des erreurs, l'alignement des lectures corrigées, le raffinement de la structure de ces derniers et la détermination des isoformes. La sélection des lectures utilise la séquence

du gène d'intérêt seul ou le génome complet, puis, elles se basent sur des seuils de couverture pour filtrer les faux positifs. Les lectures sélectionnées peuvent ensuite être corrigées avec l'outil isONcorrect pour réduire les erreurs de séquençage, augmentant ainsi la fiabilité des alignements. Un pré-filtrage des structures prédites est appliqué pour éliminer les lectures mal alignées ou hétérogène, utilisant l'alignement de exonerate. Le raffinement des alignements commence par la construction d'une matrice binaire pour représenter les alignements, et est suivi d'un pré-filtrage et d'un filtrage des points de jonction pour améliorer la cohérence des données. Les erreurs d'alignement, identifiées grâce à la structuration en bloc, sont corrigées par exonerate en mode alignement local, et le lissage des bordures consolide les jonctions exoniques en éliminant les petites erreurs de positionnement. Les lectures corrigées et alignées sont regroupées en isoformes basés sur leurs structures introniques, permettant de garantir que les isoformes prédits reflètent fidèlement la diversité transcriptomique. RNA-tailor produit plusieurs types de sorties : un fichier XSLX pour la visualisation détaillée des structures d'alignements des lectures, un fichier JSON pour un résumé structuré des données, et deux fichiers GTF pour les prédictions d'isoformes, un pour les isoformes solides et un autre pour les isoformes fragiles, facilitant ainsi l'analyse approfondie et la validation des résultats. Pour aller plus loin dans l'affinage des isoformes prédits, RNA-tailor dispose d'un module de post-traitement des isoformes. La sélection peut être filtrée selon le principe de l'inclusion des structures introniques les unes dans les autres, avec prise en compte ou non des ORF et UTR, ou par regroupement par ORF identique. Ainsi par ces différentes méthodes RNA-tailor répond aux challenges posés par la nature des lectures longues. La mauvaise qualité des séquences peut être corrigée au niveau de sa séquence directement puis au niveau de son alignement pour retrouver le bon isoforme. La faible profondeur de séquençage est compensée par sa grande sensibilité, il suffit de deux lectures pour exporter un isoformes solides. L'incomplétude des lectures est prise en compte par le réalignement des lectures, l'auto-correction mais aussi en post-traitement l'inclusion.

Valorisation et diffusion du travail réalisé depuis le début de ma thèse :

- Seqbim 2023 (présentation) : présentation
- ISMB 2023 poster and présentation virtuelle : poster / présentation virtuelle
- Git du projet, branche principale : RNA-tailor
- Git du projet, branche reproduction des résultats : benchmark

Chapitre 4

Résultats

Sommaire

Introduction

Dans cette section, nous évaluons dans un premier temps l'efficacité des différentes stratégies de sélection et de filtrage ainsi que les problématiques qu'elles soulèvent dans le cadre de l'étude des isoformes alternatifs d'épissage.

Nous évaluons la qualité des prédictions de RNA-tailor en les comparant à celles d'autres outils. Pour cela, nous utilisons plusieurs métriques proposées par l'outil GffCompare. Nous commencerons par analyser les performances de l'aligneur exonerate, un outil dont l'intégration représente une des innovations majeures de RNA-tailor, et nous les comparerons à celles de minimap2. Ensuite, nous examinerons les prédictions des outils à partir de données simulées, incluant ou non des événements de splicing artificiel. Nous terminerons par l'évaluation de ces différents outils sur des données réelles. La comparaison de la qualité des prédictions entre différents outils n'est pas simple, particulièrement lorsqu'il s'agit de données réelles sans référence absolue. Nous proposons donc des méthodes de comparaison différentes pour les études sur données réelles et simulées. Pour

les données réelles, nous choisissons d'utiliser des données issues du séquençage de 2ème génération afin de valider les jonctions d'épissage prédites par les outils. Enfin, nous nous intéresserons à l'intérêt d'une démarche exploratoire en portant une attention particulière à l'impact de chaque méthode de raffinement sur les résultats prédits. Nous verrons comment ces résultats peuvent être abordés via les fonctionnalités développées dans le module de post-traitement.

4.1 Jeux de données, simulation et évaluation

Dans cette section, nous présentons notre méthode de comparaison des résultats de prédiction des différents outils. Celle-ci a été réalisée à partir de données réelles et de données simulées. Nous justifions dans un premiers temps, le choix du simulateur de lecture, puis explicitons la méthode d'intégration d'évènements d'épissage pour produire des transcrits alternatifs synthétiques. Enfin, nous discuterons de la composition du jeux de données utilisés et du plan d'expérience. Nous terminons par une explication des métriques utilisés pour comparer les prédictions.

4.1.1 Méthodes de simulation de lectures

Avec le déploiement et l'amélioration des technologies de séquençage de troisième génération, de nombreux outils dédiés à l'analyse de ces données ont été développés. Les bonnes pratiques liées à l'amélioration continue et à la publication de ces nouveaux outils requiert de tester ces méthodes d'analyses sur des données contrôlées pour réaliser des benchmarks fiables. Cependant, la génération de telles données n'est pas toujours possible parce qu'elle engendrerait des designs d'expériences trop complexes et/ou trop onéreux. De plus, l'identification d'une vérité fiable n'est pas toujours possible. Pour pallier ce problème, plusieurs outils de simulation de lectures longues ont vu le jour. Ces derniers proposent, à partir d'un ensemble de gènes, de leurs profils d'expression donnés et de leurs séquences de référence, de modéliser des lectures en suivant un profil d'erreur caractéristique d'une technologie particulière. Dans le cadre de la simulation des longues lectures, il s'agit d'imiter les profils d'erreurs de séquençage des technologies d'Oxford Nanopore et de Pacific Biosciences, selon leurs différents modèles et mises à jour.

On s'intéresse ici à deux outils de simulation de lectures longues parmi d'autres, que nous avons sélectionnés pour leur popularité et leur facilité d'utilisation : il s'agit de NanoSim [Yan+17] et de PBSIM3 [OHA22]. Tous les deux sont capables de simuler des lectures des deux technologies à lectures longues. Néanmoins, nous nous sommes focalisés sur la simulation de jeux de données Nanopore.

NanoSim

NanoSim est un simulateur de séquences de lectures longues issues du séquençage d'Oxford Nanopore Technologies (ONT) MinION. Il est conçu pour reproduire les caractéristiques spécifiques des données ONT. Il permet de générer des lectures issues du séquençage d'ADN (NanoSim), du transcriptome (Trans-NanoSim [Haf+20]) et de méta-genome (Meta-NanoSim [Yan+23]). Dans notre cas, on s'intéresse à la simulation du transcriptome par Trans-NanoSim. Le module de Trans-NanoSim génère des lectures de séquençage qui simulent les erreurs typiques introduites par les différentes méthodes de préparation des librairies et algorithmes de base-calling. Il fonctionne en deux étapes principales. La première étape est l'apprentissage du modèle d'erreur à partir de lectures expérimentales (ou si on ne dispose pas de telles lectures, les auteurs mettent à disposition des modèles pré-entrainés). Les lectures expérimentales sont alignées contre le transcriptome de référence pour à la fois identifier leur transcrit source et caractériser le profil d'erreur des lectures. Chaque lecture simulée est associée à un transcrit source puis la séquence de l'isoforme est extraite de la référence et modifiée selon le modèle d'erreur choisi. D'après les auteurs,

même si la longueur théorique des lectures simulées devrait être la même que celle des lectures expérimentales, la longueur simulée observée est le résultat de l'apprentissage sur des données ayant des artefacts de séquençage. Un pipeline de quantification de l'abondance des transcrits est également intégré et est basé sur minimap2 pour estimer les niveaux d'expression des transcrits mais l'utilisateur peut fournir un profil d'expression (ce que nous avons fait). La caractérisation du profil d'erreurs des lectures longues se fait grâce a un modèle statistique. Pour le transcriptome, le même modèle est utilisé pour chaque type de technologie séquençage, cDNA ou RNA direct. Dans les séquences, la longueur des insertions/délétions et des *mismatch* est décidée à partir d'une loi de distribution statistique géométrique de Weibull ou de loi poisson respectivement. La probabilité de transition entre deux bases consécutives ayant des erreurs ou non est modélisée par une chaîne de Markov.

PBSIM3

PBSIM3 est un outil populaire pour la simulation de séquences longues issues des technologies de séquençage PacBio et Oxford Nanopore Technologies (ONT). Ce simulateur propose des modèles de simulation d'erreur pour différentes versions de ces technologies de séquençage. Les modèles d'erreur sont générés à partir d'un modèle HMM (Hidden Markov Model) avec critère de factorisation (FIC-HMM). Ils sont entraînés à partir de données d'alignements locaux entre des lectures longues experimentale et le transcriptome de référence, dans le cadre de l'entrainement de modèle de simulation du transcriptome. Ces alignements ont été réalisés avec l'outil LAST [FWH10]. D'après les auteurs, L'utilisation de son algorithme d'alignement local permet d'obtenir un alignement plus précis que les mappeurs comme minimap2. Toujours d'après les auteurs, cette méthode permet l'apprentissage sur une base de données plus précises et conduit à la construction de meilleurs modèles, capables de mieux capturer de la distribution des erreurs dans les lectures, y compris leur tendance à la distribution non-uniforme. Ces modèles, ne reposant pas sur un profil de $k-$mère, sont indépendants des séquences et permettent de simuler des erreurs telles que les substitutions, insertions et délétions basées sur une approche probabiliste de transition entre différents états.

Lors de la simulation d'un séquençage transcriptomique, PBSIM3 utilise un modèle basé sur la distribution de Pareto pour simuler les dégradations de début de lecture. Les auteurs soulignent que cette approche permet de reproduire fidèlement la distribution des longueurs des lectures réelles.

La performance de PBSIM3 a été comparée à d'autres simulateurs comme Badread [Wic19] et NanoSim, montrant une plus haute précision dans la simulation des modèles d'erreur, la non-uniformité des erreurs et les biais en homopolymères.

Motivations quant au choix de PBSIM3

Le choix d'utiliser PBSIM3 plutôt que NanoSim pour notre étude a été guidé par une analyse comparative des distributions des longueurs des lectures des données réelles à celles simulées. Comme illustré dans la figure 4.1, la distribution des longueurs de séquences générées par PBSIM3 correspondait davantage.

En outre, les résultats du benchmark publiés par PBSIM3 [OHA22] démontrent des performances supérieures en termes de prédiction par rapport à NanoSim, ce qui a renforcé notre choix pour les simulations dans le cadre de la thèse. Enfin, les expériences réalisées dans le cadre de cette thèse tirent parti de la fonctionnalité de PBSIM3 de pouvoir sélectionner le taux d'erreur des données simulées.

4.1.2 Simulation d'évènements d'épissage aléatoires

Afin d'évaluer la capacité des méthodes à détecter les événements d'épissage alternatif, nous avons généré des transcrits synthétiques par l'introduction d'événements d'épissage

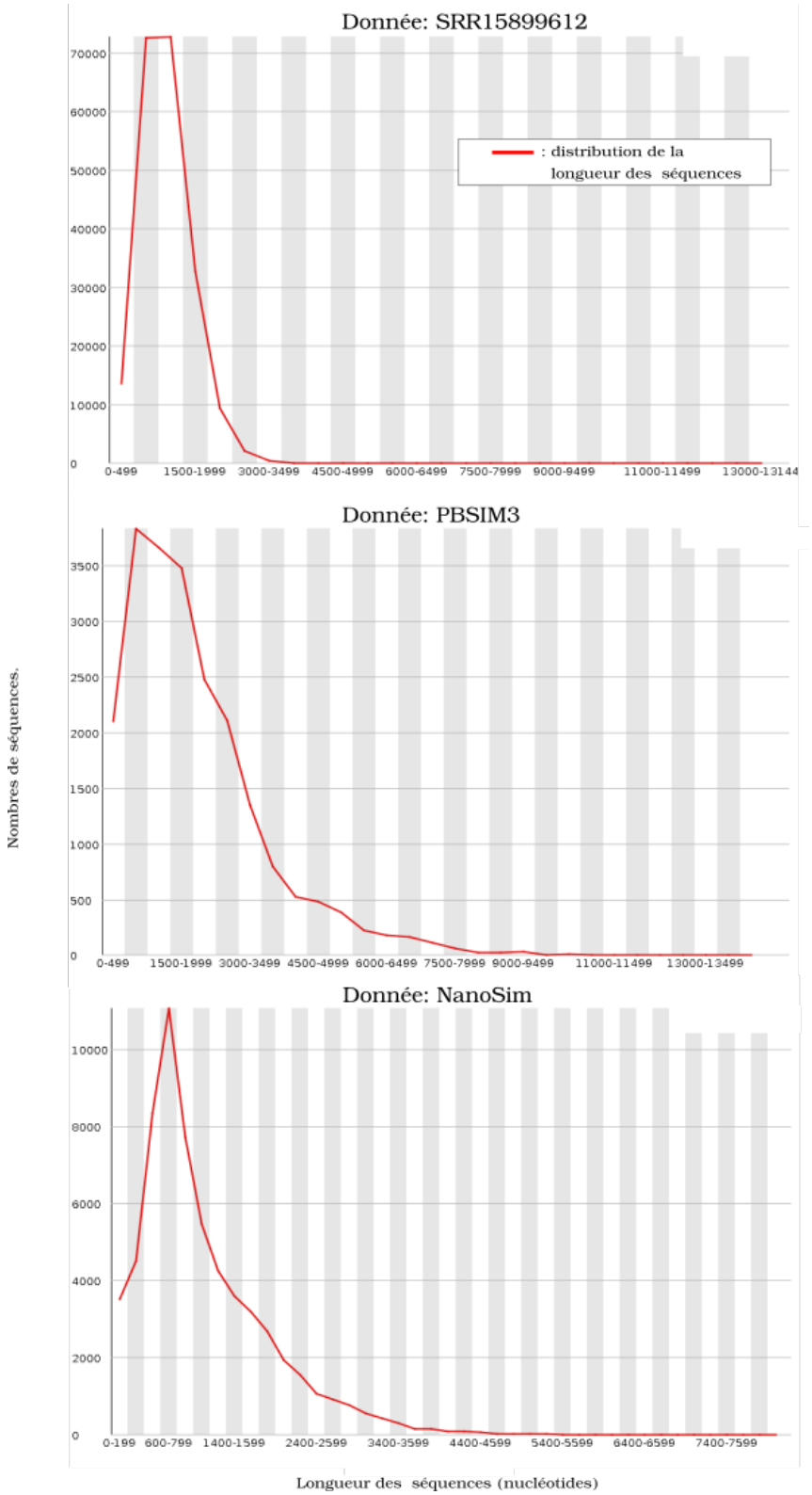


FIGURE 4.1 – Comparaison de la distribution des longueurs des lectures longues réelles et simulées. Les lectures réelles proviennent du jeu SRR15899612. Les lectures simulées l'ont été avec PBSIM3 et NanoSim avec un profil d'expression uniforme pour 280 gènes sélectionnés (voir Section 4.1.3). On s'intéresse ici au profil de la distribution plutôt qu'aux nombres de séquences en ordonnées. Les distributions des longueurs des lectures réelles et des lectures simulées par PBSIM3 ont un profil similaire, avec un pic du nombre de lectures de longueur comprise entre 750 et 1200 nucléotides. A l'inverse, les lectures simulées par NanoSim possèdent un pic étroit en 750 nucléotides uniquement. Le profil de distribution proposé par PBSIM3 nous a donc semblé plus réaliste.

aléatoires dans des isoformes connus. Nous avons choisi de simuler les trois types de modifications d'épissage les plus courants : exon cassette (EC), rétention d'intron (IR), et début/fin d'exon alternatif en tandem (ASS). La simulation des événements est réalisée par gène. Le nombre de nouveaux transcrits synthétiques est choisi aléatoirement en fonction du nombre de transcrits connus pour chaque gène : il varie de zéro à la moitié du nombre de transcrits connus. Chaque nouvel isoforme ne peut avoir qu'un seul type de modification parmi les trois. Les règles de création de chacun de ces événements sont les suivantes :

- le saut d'exon est simulé par la suppression aléatoire d'un ou plusieurs exons internes allant d'un seul exon jusqu'à un tiers du nombre total d'exons ;
- la rétention d'intron artificielle est réalisée par la fusion de deux exons consécutifs. Un exon est tiré au sort aléatoirement, à l'exclusion du premier et du dernier exon, et il est fusionné avec l'exon suivant ;
- la simulation d'un ASS consiste à rechercher des sites d'épissage en tandem dans un rayon de 12 nucléotides en amont et en aval de chaque exon interne de chaque transcrit.

4.1.3 Jeux de données

Jeu de données humain

Les données concernant l'homme sont extraites de la publication de Freddie. On dispose, à partir de la même expérience de séquençage, de lectures longues séquencées grâce a la technologie Oxford Nanopore PromethION (numéros d'accession SRR15899612) et de lectures courtes générées à partir de la technologie Illumina (numéros d'accession SRR15899613). Les lectures courtes permettront de valider des sites d'épissage prédites par les lectures longues (voir légende figure 4.12). D'après les résultats d'alignement en sortie de minimap2, le taux d'erreur moyen de séquençage pour les lectures longues est autour de 8,4 %. Les auteurs de Freddie ont sélectionné 294 gènes à analyser connus pour contenir des isoformes alternatifs d'épissage exprimés dans les conditions de ces expériences de séquençage. Tous les gènes sont multi-exoniques. Cependant, nous avons ajouté un niveau de filtrage supplémentaire basé sur les qualificatifs des annotations Ensembl. L'attribut "tag" de chaque transcrit propose une information supplémentaire sur les transcrits. À partir des 294 gènes initiaux, on garde ainsi 280 gènes produisant au moins une protéine multi-exonique. Ces 280 gènes forment un total de 1153 transcrits possédant le tag "protein coding". Parmi eux 875 transcrits codent pour des protéine multi-exonique viable, c'est-à-dire possèdant l'un des tags "basic, CCDS, Ensembl_Cannonical, MANE_Plus_Clinical, MANE_Select" (les tags exclus sont "mRNA_end_NF, mRNA_start_NF, cds_end_NF, cds_start_NF, seleno"). Le tableau 4.1 résume ces données.

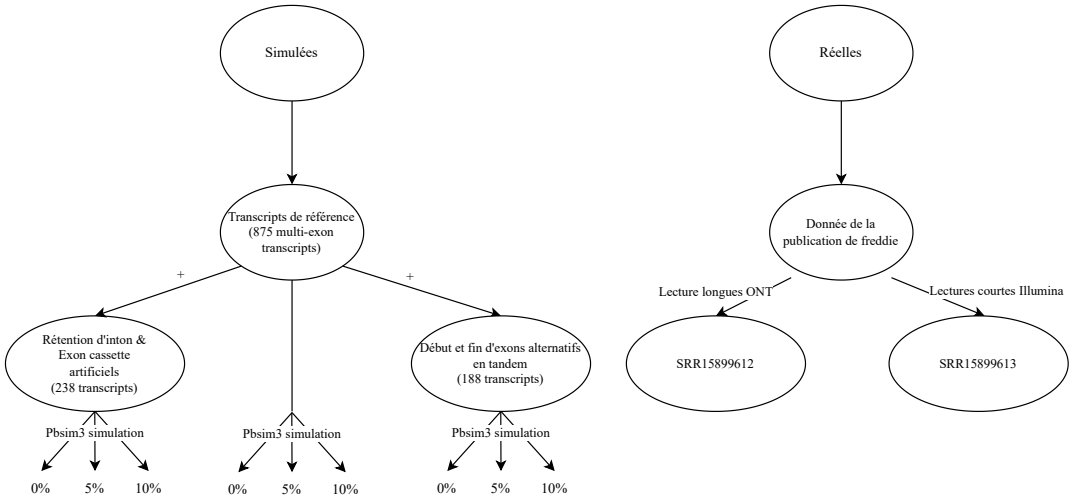
À partir du même jeu de gènes, nous avons construit plusieurs jeux de données simulées, sur la base des transcrits annotés. Pour éviter de générer des données contenant trop de transcrits synthétiques, ce qui éloignerait davantage les données analysées de la réalité, nous choisissons de séparer l'expérience en deux sous-expériences. La première simule l'occurrence de rétentions d'intron et d'exons cassette et la seconde l'apparition de début et fin alternatifs d'exon. Plus précisément les événements d'épissage simulés sont consignés dans deux fichiers GTF de transcrits modifiés : un pour les ASS et un pour les EC et les IR. Chaque fichier est ensuite combiné avec le fichier GTF des 875 transcrits de référence. Les fichiers GTF résultants sont fournis comme entrée à PBSIM3 pour simuler les lectures. Au total 426 transcrits modifiés ont été ajoutés aux transcrits de références. Le nombre de modifications effectuées dans nos expériences est résumé dans le tableau 4.2. Les capacités de prédictions ont été évaluées à trois niveaux d'erreurs de séquençage différentes (0%, 5%, 10%) pour analyser son influence sur la qualité des prédictions. Un résumé du plan de l'expérience est visible en figure 4.2.

TABLE 4.1 – Étude de l'unicité et de l'inclusion des structures isoformes des transcrits conservés pour les simulations pour les 280 gènes. Etude de l'unicité et de l'inclusion des structures isoformes connues les unes par rapport aux autres selon différents filtres de sélection des isoformes. Les tags "Protein Coding" et "transcrit Coding" sont respectivement inclus dans les attributs gene_biotype et transcrit_biotype. L'attribut "tag" de chaque transcrit propose une information supplémentaire sur le transcrit. Les transcrits dont l'attribut "tag" contient l'une des annotations suivantes sont conservés : "basic, CCDS, Ensembl_Cannonical, MANE_Plus_Clinical, MANE_Select".

Tags appliqués sur les transcrits	nombre d'isoformes	nombre de structures introniques uniques	nombre de structures introniques uniques non incluses
Protein Coding only	2210	2186	1804
Protein Coding + transcript coding	1153	1139	1051
Protein Coding + transcript coding + complete CDS	875	868	757

TABLE 4.2 – Table du nombre de transcrits modifiés par modification à partir de 875 isoformes de référence.

		Genes	Isoformes	Lectures
SRR15899612		294		203844
Lectures simulées	Connus	280	875	20× le nombre d'isoformes
	EC		118	
	IR		120	
	ASS		188	



Expérience	Nombre de lectures	Taux d'erreur	Nombre de gènes	Nombre de transcrits
SRR15899612	203844	$\approx 8,4\%$	294	
PBSIM_0 PBSIM_5 PBSIM_10	875×20	0% 5% 10%	280	875
PBSIM_IREC_0 PBSIM_IREC_5 PBSIM_IREC_10	1113×20	0% 5% 10%	280	1113
PBSIM_TASS_0 PBSIM_TASS_5 PBSIM_TASS_10	1063×20	0% 5% 10%	280	1063

FIGURE 4.2 – Schéma récapitulatif des jeux de données pour l'humain utilisés dans les différentes expériences d'évaluation des performances de prédictions des différents outils d'identification des isoformes alternatifs d'épissage. La composition en isoformes des deux expériences de simulées est réalisée en additionnant les transcrits connus et les transcrits issus de la génération d'événement aléatoires, formant une base de 1113 transcrits pour l'étude des ES et IR, et une base de 1063 transcrits pour l'analyse des évènements de type ASS.

Autre jeux de données

Mouse brain cDNA Ce jeu de données disponible sur SRA sous l'accession PRJEB25574 a été généré par le Genoscope dans le cadre de l'ANR ASTER. Il s'agit de données de lectures longues cDNA séquencées par un séquenceur MinION d'Oxford Nanopore avec une flowcell r9.4. Le gène « ENSMUSG00000000827 » est exprimé dans ce jeux de données. Son analyse a permis de produire les données montrées en figure 4.16.

4.1.4 Méthodes d'évaluation des résultats de prédictions

Durant le temps de la thèse, j'ai utilisé deux méthodes d'évaluation différentes des résultats de prédictions publiées et largement utilisées : GffCompare [PP20] et SQANTI3 [Par+24].

Les deux outils proposent une évaluation des prédictions en comparant deux fichiers GTF, dont l'un joue le rôle de base de vérité et l'autre jour le rôle de prédiction. Pour ce faire, ils classent chaque transcrit prédit selon une classification propre (voir figure 4.3) mais qui présentent tout de même des similitudes.

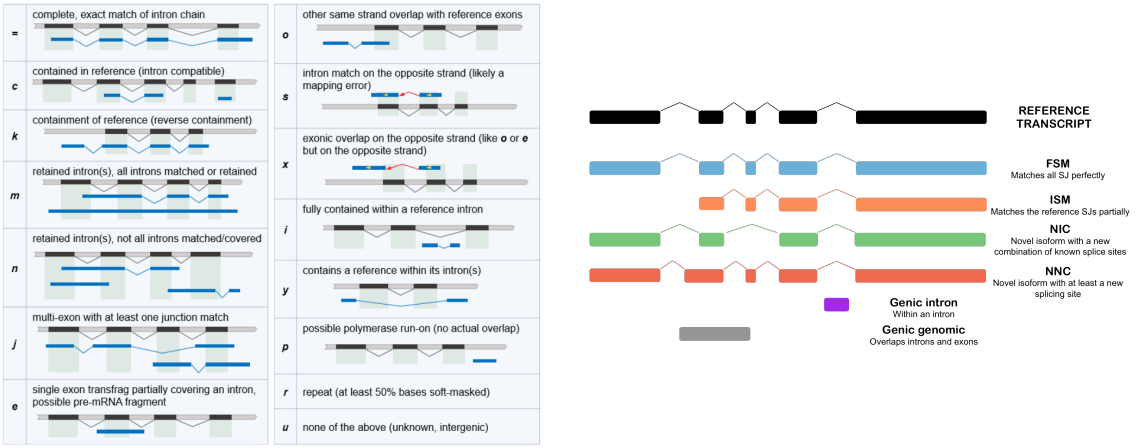


FIGURE 4.3 – Classification proposée par GffCompare (à gauche) et classification proposée par SQANTI3 (à droite). Pour GffCompare les lettres correspondent au code retrouvé dans le fichier résultat d'analyse de GffCompare. Tirés de [PP20] et [Par+24].

SQANTI3 propose de classer les isoformes en 6 catégories :

- FSM (Full Splice Match) : toutes les jonctions (tous les introns) correspondent entre l'isoforme de référence et l'isoforme prédit (les positions du début du premier exon et de la fin de dernier exon ne sont pas prises en compte) ;
- ISM (Incomplete Splice Match) : toutes les jonctions de l'isoforme prédit correspondent à la référence mais il peut manquer des exons en 5' et en 3' ;
- NIC (Novel In Catalog) : l'isoforme prédit utilise une combinaison de jonctions connues (mais pas nécessairement une combinaison d'introns connus) ;
- NNC (Novel Not in Catalog) : l'isoforme prédit utilise une ou plusieurs jonctions non connues ;
- Genic Intron : l'isoforme prédit est entièrement contenu dans un intron connu ;
- Genic Genomic : l'isoforme prédit chevauche des introns et des exons connus.

Des sous-catégories viennent compléter la classification. On note également deux catégories supplémentaires permettant de classer les isoformes prédits, pour le cas où l'isoforme est prédit dans une région intergénique ou en antisens.

GffCompare offre une classification détaillée des possibilités structurelles des isoformes en comparaison à une référence, réparties en 15 catégories. On ne cite ici que quelques-unes :

- = : correspond à FSM ;
- c : correspond à ISM ;

— k : permet d'identifier des isoformes prédits dont une référence est un ISM vis-à-vis du prédit ;
— i : correspond à Genic Intron ;

Il permet d'identifier différents types de rétention d'introns, qu'elle soit complète ou partielle. Comme SQANTI3, certains transcrits sont également identifiés comme des erreurs de mapping lorsque le transcrit correspond à une référence mais dans le mauvais sens d'alignement (x), et également les transcrits mono-exonique qui sont contenus ou se superposent avec une référence, ainsi que ceux intergéniques. En plus de la classification GffCompare permet d'obtenir des statistiques sur les exons prédits (précision et sensibilité), les introns prédits (précision et sensibilité), ainsi que sur les jonctions prédites (nommées *base* dans les fichiers de sortie).

Les deux classifications proposées par SQANTI3 and GffCompare reprennent des similarités avec des appellations différentes. On retrouve dans les deux cas, une définition des transcrits FSM identique. Celle-ci est basée sur le *match* exact de la totalité de la chaîne intronique entre un isoforme de référence et un transcrit prédit. Cette similarité se retrouve dans la définition commune des ISM, représentant un transcrit issu d'un match exact mais incomplet des jonctions introniques. Cette définition de match exact basé sur les jonctions intronique est communément adopté pour étudier les longs reads. Elle permet de tenir compte du fait de la dégradation en 5' et 3' des lectures lors du processus de séquençage, qui mène à une imprécision dans la détermination des TTS et TSS.

Nous avons finalement opté pour l'utilisation de GffCompare.

4.2 Évaluation de l'efficacité des méthodes de sélection et de filtrage des lectures du pipeline de RNA-tailor

La sélection des lectures à partir de la base de données est une première étape cruciale pour la prédiction des isoformes. Ici, on étudie la qualité de la sélection des lectures selon les deux méthodes proposées par RNA-tailor : soit le génome entier avec minimap2, soit la séquence génomique du gène avec megablast. L'évaluation de l'efficacité de l'étape de filtrage du pipeline de RNA-tailor est analysée sur la sélection des lectures à partir des données simulées PBSIM_10. Enfin, on aborde l'effet du filtre sur les introns uniques à partir de la distribution des structures introniques en sortie d'exonerate.

Sélection des lectures sur les données réelles

Le nombre de lectures sélectionnées par l'une et l'autre méthode sur les données réelles (jeu de données SRR15899612) est présenté dans la figure 4.4. On s'aperçoit que l'utilisation

Nombre de lectures	Spécifiques	Communes	Total
megablast	1739	154375	156114
minimap2	26445		180820

FIGURE 4.4 – Comparaison du nombre total de lectures sélectionnées pour l'analyse des 280 gènes à partir du jeu de lectures SRR15899612 (203844 lectures au total), à partir de la séquence du gène seulement avec megablast, ou à partir du génome entier avec minimap2.

de minimap2 sur le génome entier permet de capturer un plus grand nombre de lectures par rapport à l'utilisation de megablast sur seulement la séquence du gène. On peut expliquer cette différence par les bonnes performances de minimap2, développé spécifiquement pour l'alignement de lectures longues ayant un fort taux d'erreurs. L'information supplémentaire donnée par le génome entier met en compétition des sites d'alignement potentiels ce qui facilite le choix des alignements pertinents. Ce second point permet d'expliquer la sélection de lectures par megablast qui ne sont pas sélectionnées par minimap2. C'est notamment un levier important pour éviter des contaminations de lectures provenant de l'expression

d'autres gènes. Le risque de contamination est encore plus important pour les gènes possédant des gènes paralogues qui peuvent exprimer des transcrits susceptibles à partir de la séquence de chacun des gènes paralogues. On observe également que la sélection la grande majorité des lectures sélectionnées le sont par les deux méthodes (respectivement 85% et 98% pour minimap2 et megablast). Cependant, utiliser la référence du génome entier avec minimap2 permet de sélectionner 17% de lectures supplémentaires. On peut se demander si les lectures supplémentaires sélectionnées peuvent permettre à RNA-tailor de prédire des isoformes connus supplémentaires.

On étudie cette question en comparant les prédictions de RNA-tailor en fonction des deux méthodes de sélection et en comparaison avec une base de vérité. La base de vérité est ici construite à partir des annotations des 280 gènes. On sélectionne les isoformes dont toutes les jonctions sont soutenues par les lectures courtes du jeu de données SRR15899613, soit au total 562 isoformes.

On commence par s'intéresser uniquement aux isoformes solides prédits par RNA-tailor (voir définition 17) soutenus par au moins une lecture spécifique. Les résultats sont présentés Figure 4.5. Même si le nombre d'isoformes connus soutenus par les lectures est

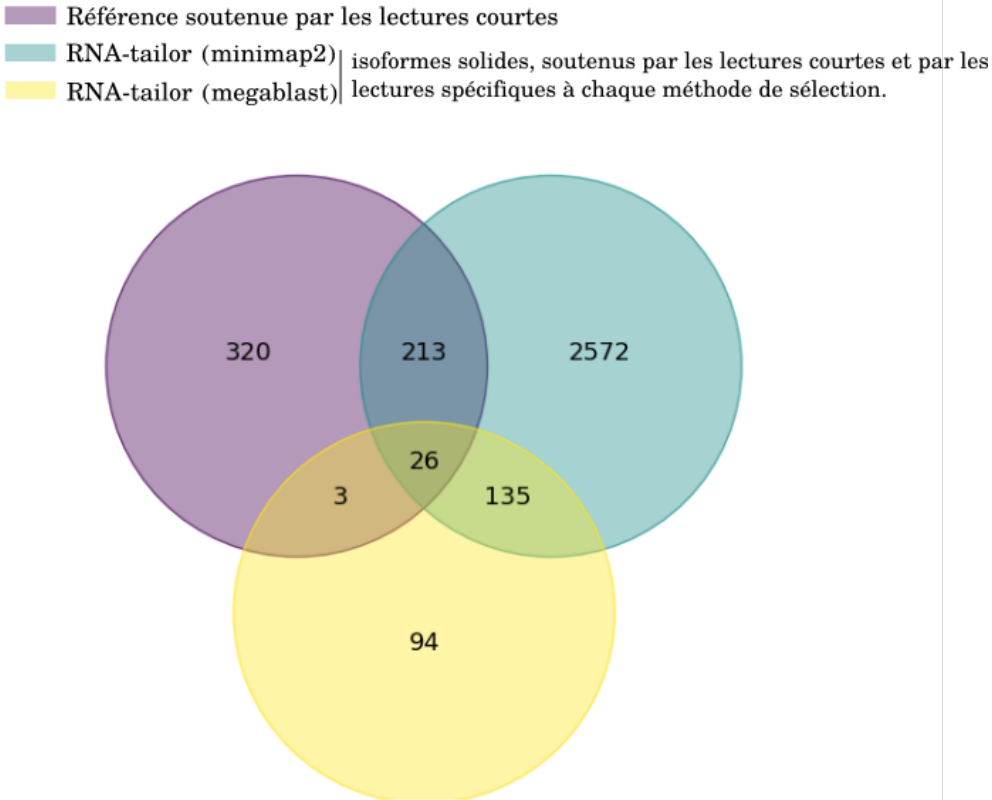


FIGURE 4.5 – Comparaison des prédictions d'isoformes solides de RNA-tailor impliquant des lectures spécifiques à chacune des méthodes. Pour megablast RNA-tailor prédit 258 $(3+26+135+94)$isoformes impliquant l'une des 1739 lectures spécifiques. Pour minimap2 RNA-tailor prédit 2946 $(213+26+135+2572)$ isoformes solides impliquant l'une des 26445 lectures spécifiques.

plus grand pour minimap2 ($239 = 213 + 26$) que pour megablast ($29 = 26 + 3$), la part des lectures sélectionnées soutenant un isoforme connu est similaire (11,1% pour megablast et 8,1% pour minimap2). La sélection plus abondante de minimap2 permet donc de soutenir plus d'isoformes connus au total mais soutient également plus de nouvelles structures. Dans le cadre de l'étude des lectures longues, on peut préconiser l'utilisation de minimap2 pour sa plus grande sensibilité comme la profondeur de séquençage peut être un problème limitant.

On poursuit maintenant en s'intéressant à l'ensemble des isoformes solides prédits par RNA-tailor. Les résultats sont présentés Figure 4.6. On peut apprécier les différences de

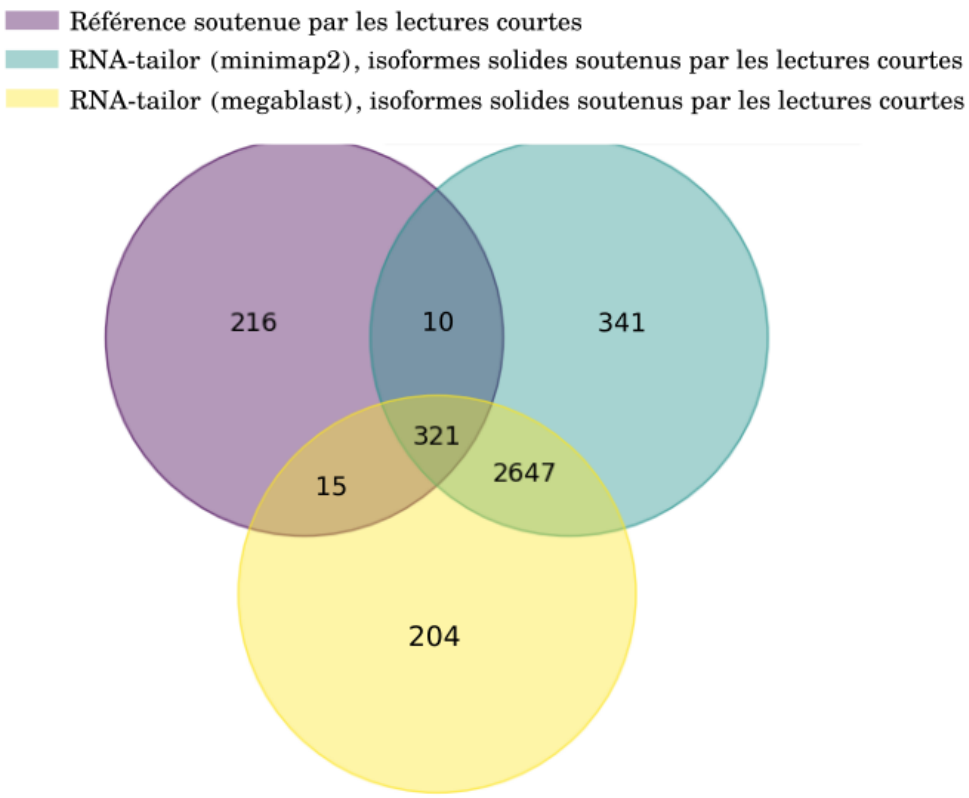


FIGURE 4.6 – Comparaison de toutes les prédictions d'isoformes solides de RNA-tailor. Pour megablast RNA-tailor prédit 3319 ($10 + 321 + 2647 + 341$) isoformes. Pour minimap2 RNA-tailor prédit 3187 ($15 + 321 + 2647 + 204$) isoformes. Parmi les 562 isoformes connus soutenus par les lectures courtes, 10 isoformes prédits sont spécifiques à la sélection via megablast tandis que 15 isoformes prédits sont spécifiques à la sélection via minimap2.

composition entre les isoformes prédits par la sélection à partir du génome entier et ceux prédits à partir la séquence du gène seulement. On s'aperçoit que la majorité des prédictions sont similaires respectivement 89% et 93% pour l'échelle du gène et du génome. Cela reste cohérent avec les résultats de comparaison de sélection des lectures qui étaient en grande partie identiques. Les différences restantes représentent respectivement 10% et 4% des prédictions. Parmi les 562 isoformes connus soutenus par les lectures courtes, seuls

10 isoformes prédits sont spécifiques à la sélection via megablast tandis que 15 isoformes prédits sont spécifiques à la sélection via minimap2, soit moins de 0.4% du total. Ainsi, les lectures supplémentaires sélectionnées par chacun des outils ont pu apporter une information minime mais pertinente et différente. La sélection d'un plus grand nombre de lectures grâce au génome entier a donc permise d'augmenter le support des isoformes plutôt que la production de nouvelles structures. Ainsi même si la grande majorité du signal est capturée par les deux méthodes, il reste préférable d'utiliser le génome entier pour la sélection. L'étude de la composition en lectures des 10 et 15 isoformes connus prédits spécifiquement par chacune des deux méthodes montre que ces isoformes prédits spécifiques sont bien soutenus par des lectures complètement distinctes (1171 lectures spécifiques pour megablast et 385 lectures spécifiques à minimap2). La prédiction différentielle des isoformes qui en découle est donc due à la sélection différentielle des lectures. Ici, chaque méthode permet d'apporter une information pertinente et complémentaire.

On termine avec une analyse identique à la précédente concernant les isoformes fragiles prédits par RNA-tailor. Les résultats sont présentés Figure 4.7. On s'aperçoit que la

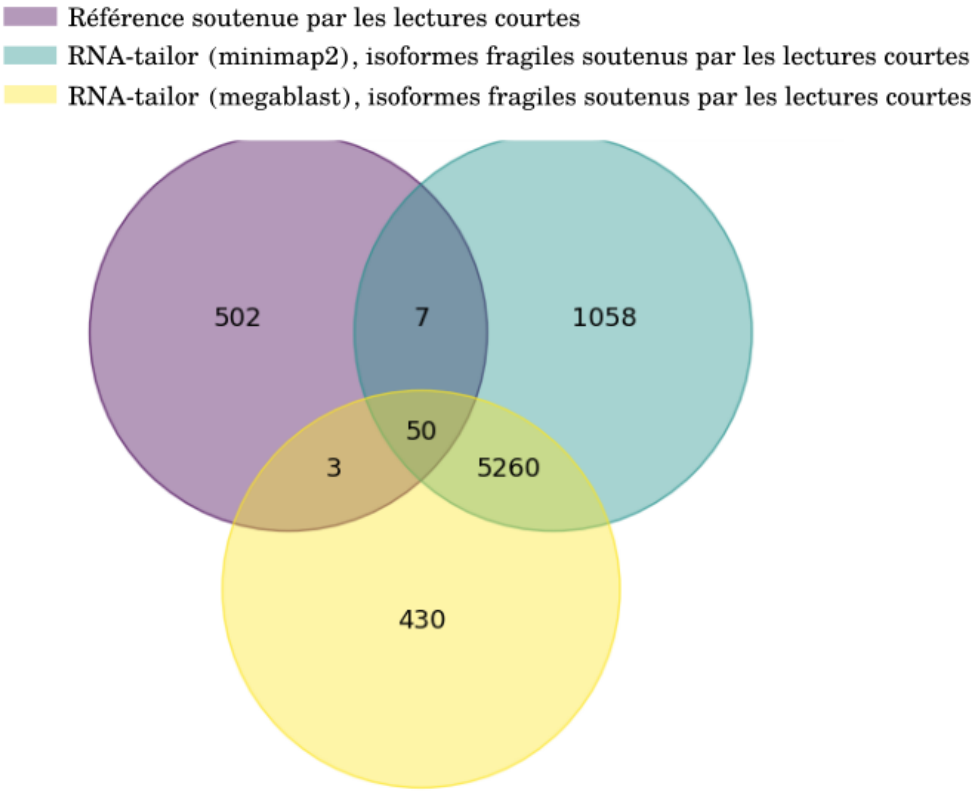


FIGURE 4.7 – Comparaison de toutes les prédictions d'isoformes fragiles de RNA-tailor.

distribution des isoformes prédits suit le même tendance que pour les isoformes solides.

Sélection sur les données simulées et influence du filtrage

Pour évaluer la part de contamination (c'est-à-dire la part des lectures sélectionnées pour un gène alors qu'elles proviennent de la simulation des transcripts d'un autre gène) dans les lectures sélectionnées par megablast et minimap2 pour chaque gène, on utilise des lectures simulées. Pour celles-ci, on connaît leurs gènes d'origine et on peut vérifier qu'elles sont correctement sélectionnées par chacune des méthodes. Le risque est que pour la méthode de sélection par référence du gène seul induise la sélection de lecture non issue du gène si sa séquence s'aligne avec une séquence intronique ou s'ils sont paralogues. On étudie dans un premier temps la composition en lectures sélectionnées par chaque méthode dans la figure 4.8a. On retrouve la tendance observée dans les données réelles, avec minimap2

Nombre de lectures	Spécifiques	Communes	Total
megablast	1	13437	13438
minimap2	3755		17192

a) avant filtrage

Nombre de lectures	Spécifiques	Communes	Total
megablast	62	13284	13346
minimap2	1623		14907

b) après filtrage

FIGURE 4.8 – Comparaison de la sélection des lectures à partir des deux méthodes différentes utilisant minimap2 et megablast : a) avant application du filtrage (décrit en Section 3.2.3), et b) après application du filtrage. Les données utilisées ici ont été générées à partir de 875 isoformes par PBSIM_10.

qui, de part son algorithme d'alignement adapté aux erreurs de séquençage lecture longue et la référence du génome, possède une plus grande sensibilité à la sélection des lectures.

Un point intéressant est d'observer ce qu'il en est de ces lectures spécifiques une fois l'étape de filtrage précoce effectuée. La comparaison des lectures après cette étape est présentée dans la figure 4.8b. On observe que le filtrage supprime davantage de lectures sélectionnées par minimap2 (2285) que par megablast (92), dont 153 lectures longues sélectionnées communément par les deux méthodes. Ainsi, la sélection par megablast montre une plus grande homogénéité en terme de structure d'épissage vis-à-vis de nos critères de filtrage. Pour compléter l'analyse, on s'intéresse à la part des lectures contaminantes dans la sélection effectué par megablast. La table 4.3 montre megablast sélectionne peu de séquences contaminantes. Elle ne représente que 1,5% du total des lectures avant filtrage et ce taux passe à 1% après filtrage. Parmi les 92 lectures supprimées par le filtrage dans la sélection faite par megablast, 75 lectures sont contaminantes. Ici, le filtrage permet de réduire de 36% les erreurs de sélection faites par megablast. Les critères de reconnaissance des lectures hétérogènes ont permis un filtrage précis car seules 17 (soit 18,4%) lectures ont été supprimées alors qu'elles avaient été correctement sélectionnées. Leur suppression peut être due au fait qu'il s'agissait de lectures monoexonique, provenant d'une forte dégradation de la séquence des transcripts d'origine. Notre méthode de sélection des lectures misant sur la longueur d'alignement de la lecture sur la référence laisse donc passer peu de contamination dans le cadre de notre expérience. Une partie de ces lectures contaminantes sont reconnues par notre filtre après alignement par exonerate sur la référence du gène et permet d'identifier que ces lectures sont monoexoniques ou possédant une structure intronique unique.

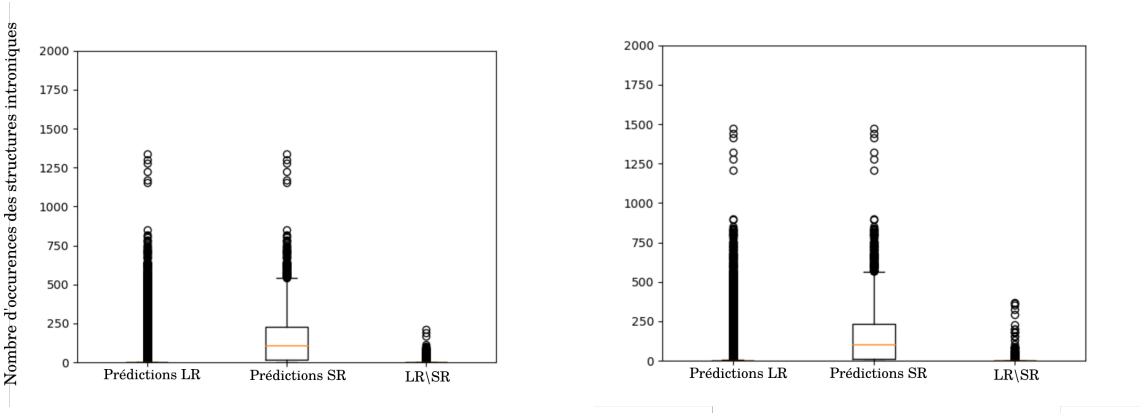
Discussion sur la motivation de l'étape de filtrage

L'étape de filtrage des lectures sélectionnées est décrite en section 3.2.3. Elle a lieu avant les étapes de raffinement. Son objectif est d'enlever précocement les lectures faussement

TABLE 4.3 – Table d'appréciation de l'effet du filtrage des lectures sur le nombre de lectures totales gardées et sur le nombre de reads comtanimants pour megablast.

	Lectures correctement sélectionnées	Erreurs de sélections
avant filtrage	13397	210
après filtrage	13231	135

sélectionnées. Ce filtrage a lieu après l'alignement des lectures non corrigées par exonerate. L'intérêt est d'assainir l'ensemble des lectures sélectionnées en enlevant des erreurs potentielles avant l'étape d'auto-correction par isONcorrect. Son effet a été observé sur l'évolution de composition en lectures dans la figure 4.8 et la table 4.3. Le filtrage concerne les lectures mono-exoniques et les structures introniques faiblement représentées. Le filtrage des structures introniques a été motivé par l'étude de la distribution des structures introniques issue des alignements splicés de exonerate que l'on compare à ceux produit pas minimap2, visible en 4.9. On utilise les alignements des lectures courtes, comme référence



a) exonerate b) minimap2

FIGURE 4.9 – Distribution de l'abondance des structures introniques prédites en sortie d'alignement des lectures longues SRR15899612 (« LR ») comparée à celle de l'alignement des lectures courtes SRR15899613 (« SR ») par a) exonerate, et b) minimap2. La troisième boîte à moustache correspond aux prédictions introniques uniquement retrouvés dans les prédictions faites par l'aligneur de lectures longues.

des structures introniques présentes dans l'expérience de séquençage. La figure montre une dispersion plus importante des introns prédits par minimap2 et non soutenus par la référence par rapport aux prédictions de exonerate. Dans ce cas, les structures non soutenues sont donc plus simplement identifiables par un seuil de support de lecture pour exonerate que minimap2. Dans les deux cas, l'application d'un seuil de support est une bonne méthode d'élimination des erreurs d'alignement.

4.3 Évaluation des prédictions d'outils d'identification de transcriptome sur des données simulées

Nous nous intéressons ici à l'analyse de la prédiction d'isoformes à partir des jeux de données PBSIM_IREC_0, PBSIM_IREC_5, PBSIM_IREC_10, PBSIM_TASS_0, PBSIM_TASS_5, PBSIM_TASS_10.

4.3.1 Comparaison des résultats d'alignements de exonerate et de minimap2

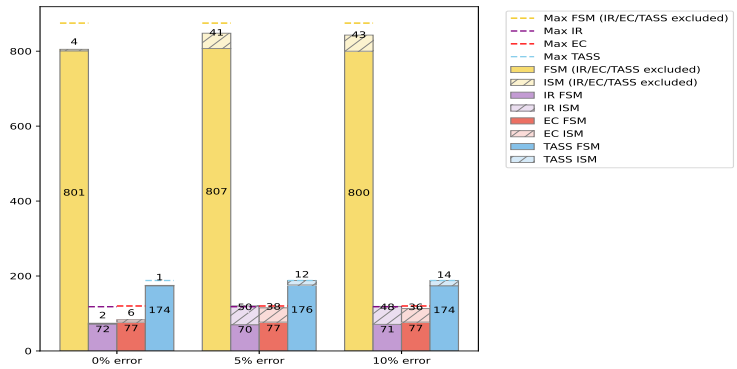
Alors que minimap2 et exonerate sont utilisés pour produire des alignements épissés en tant qu'entrées pour les méthodes de détection des isoformes, on peut se demander s'ils fournissent des alignements pertinents et pourquoi ces outils ne sont pas utilisés directement pour détecter les isoformes. La Figure 4.10 compare les performances de minimap2 et exonerate dans leur capacité à détecter les FSM et ISM sur des ensembles de données simulées, incluant ainsi des transcrits simulés ASS, IR et ES. Pour exonerate, nous comparons les résultats avec ou sans correction des lectures avec isONcorrect.

Bien qu'exonerate surpasse constamment minimap2 dans la reconnaissance des transcrits connus et l'identification des événements d'épissage alternatif simulés, minimap2 maintient des performances stables malgré les variations des taux d'erreur de séquençage. À mesure que le taux d'erreur diminue, les performances d'identification d'exonerate s'améliorent pour les ES et IR, tout en restant bonnes pour identifier les ASS. Les comportements distincts de ces outils peuvent provenir des algorithmes heuristiques dans le processus d'alignement de minimap2, qui privilégie l'alignement aux sites d'épissage canoniques. Cela peut occasionnellement conduire à des erreurs, notamment avec des sites d'épissage faibles ou non canoniques tels que NAGNAG ou GYNGYN. À l'inverse, l'alignement d'exonerate est moins guidé vers les sites d'épissage non canoniques. La correction automatique de lecture (avec isONcorrect) sur des données buitées affecte défavorablement la détection des ASS, tout en améliorant légèrement la reconnaissance des IR et ES. Cela est dû au fait que isONcorrect crée des séquences consensus avec [Vas+17]. Lorsque la génération d'un consensus est appliqué aux niveaux de site d'épissage dans un contexte où chaque gène présente un événement ASS, isONcorrect peut alors introduire des erreurs. En effet, ces événements alternatifs sont courts et risquent plus d'être gommés par la correction que des évènements longs comme les IR et ES.

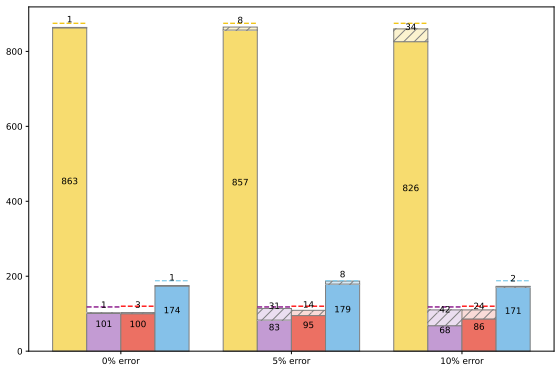
Bien qu'exonerate semble légèrement meilleur pour détecter les événements d'épissage, les deux aligneurs possèdent une sensibilité de plus de 90% pour la reconnaissance d'isoformes connues. Leur taux élevé d'ISM démontre leur capacité à capter l'ensemble du signal d'épissage. Malgré leur grande sensibilité, les alignements bruts des lectures génère un grand nombre de nouvelles structures introniques conduisant à une précision globale médiocre. Par exemple, exonerate et minimap2 ont une valeur de précision pour l'identification des FSM entre 8 et 10% pour les données simulées, tant pour les données ayant 10% qu'à 5% de taux d'erreur. Il existe donc un besoin de méthodes capables d'extraire le signal pertinent du bruit. De plus, même si la plupart des isoformes sont identifiés, certains peuvent uniquement être soutenus par une seule lecture et le manque de précision montre qu'il y a beaucoup de bruit. Ainsi, extraire des isoformes probables à partir des alignements bruts reste une tâche difficile. Par exemple, le nombre de structures introniques uniques générées par exonerate est de plus de 10k pour l'ensemble de données simulées. Néanmoins, ces aligneurs fournissent une bonne base pour les outils conçus pour identifier les isoformes d'épissage grâce à leur sensibilité. À mesure que le taux d'erreur du séquençage de longues lectures diminue grâce aux avancées technologiques, ces résultats laissent entrevoir que l'utilisation d'exonerate pourrait conduire à une meilleure identification des jonctions d'épissage.

4.3.2 Évaluation des performances de prédictions de RNA-tailor contre FLAIR et Freddie

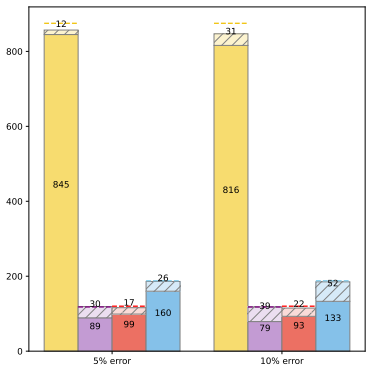
Dans cette partie, nous analysons les résultats de prédiction de RNA-tailor et FLAIR sur des jeux de données simulés. Freddie a été exclu de l'analyse car il proposait des résultats trop inférieurs à ceux de ses concurrents. En effet, afin d'identifier au moins un FSM, le développement d'un script de traitement spécifique de ses résultats a été nécessaire. La méthode d'analyse devenant différente, nous avons choisi de l'exclure du benchmark car



(a) minimap2



(b) exonerate sans correction des lectures



(c) exonerate avec correction des lectures

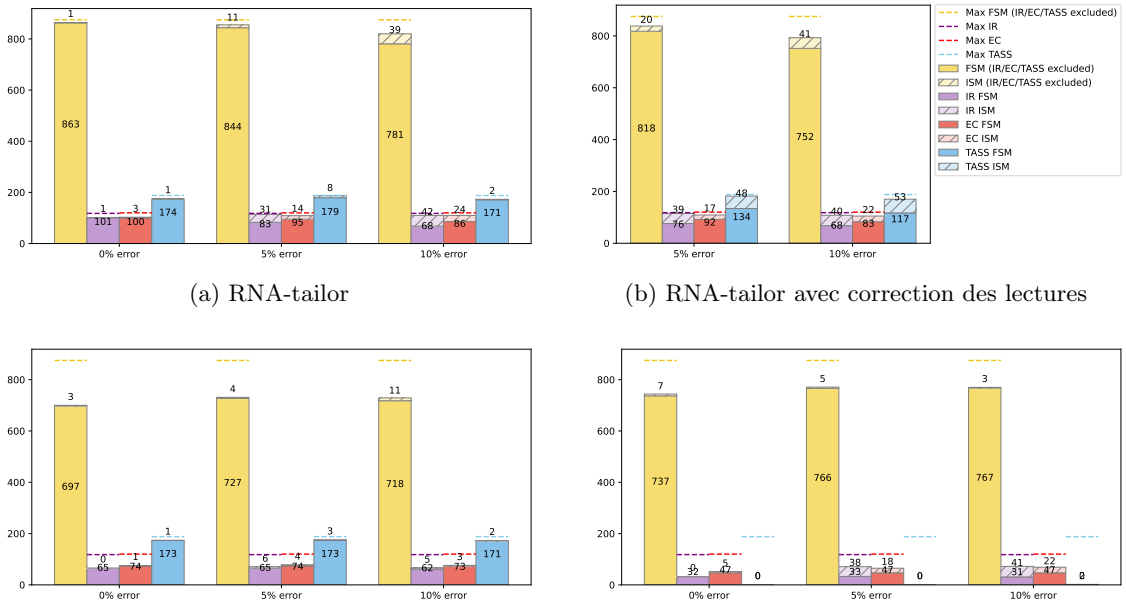
FIGURE 4.10 – Comparaison de la capacité à identifier les bons isoformes entre minimap2 et exonerate (avec et sans correction d'erreurs par isONcorrect) à partir de données simulées. Nombre de transcrits récupérés (FSM) pour chaque événement : en jaune les isoformes connues, en violet les isoformes avec un IR simulé ; en rouge les isoformes avec un ES simulé, et en bleu les isoformes avec un ASS simulé. Les barres pleines indiquent le nombre de FSM trouvés, tandis que les barres hachurées indiquent le nombre de ISM (calculé avec GffCompare). La ligne pointillée représente le nombre d'isoformes attendus pour chaque événement.

les comparaisons directes des résultats étaient faussées.

Nous utilisons ici les références des transcrits connus ou avec modifications présentés en figure 4.2 pour évaluer les performances des outils. Pour l'ensemble des transcrits, 1301 ($875 + 188 + 238$) transcrits totaux, répartis en plusieurs jeux de données, 20 lectures ont été simulées. Le fichier GTF donné en entrée à FLAIR guidé contient les 875 isoformes connus, pour se placer dans le cadre d'une exploration sans *a priori*. La Figure 4.11 montre les prédictions sur les isoformes connus (annotés) et artificiels.

Cette méthode nous permet ainsi d'apprécier la capacité de RNA-tailor et FLAIR à identifier des événements d'épissage spécifiques parmi d'autres isoformes de transcrits connus.

Identification des isoformes connus. RNA-tailor atteint le meilleur taux d'identification de FSM (barres jaunes) avec une sensibilité comprise entre 86% (avec correction, taux d'erreur de 10%) et 96% (sans correction, taux d'erreur de 5%). RNA-tailor sans correction de lecture a toujours un nombre de FSM identifié plus élevé qu'avec la correction de lecture pour les taux d'erreur à 5 et 10%. Cela peut-être dû au processus de correction, qui sélectionne un consensus par multi-alignement de plusieurs lectures. D'autre part, FLAIR guidé a retrouvé plus d'isoformes de référence en FSM que FLAIR non guidé. C'est le résultat



(a) RNA-tailor

(b) RNA-tailor avec correction des lectures

(c) FLAIR non guidé

(d) FLAIR guidé

FIGURE 4.11 – Comparaison de la capacité à trouver les isoformes simulés par RNA-tailor (avec et sans correction d'erreurs par isONcorrect) et FLAIR (versions guidée et non guidée) sur des jeux de données simulés. Le nombre d'isoformes retrouvés pour chaque événement est situé en haut de chaque bâton du diagramme : en jaune les isoformes connus, en violet les isoformes artificiels avec un IR simulé ; en rouge avec un ES simulé, et en bleu avec un ASS simulé. Les barres pleines donnent le nombre de FSM trouvés, tandis que les hachurées donnent le nombre de ISM (calculé avec GffCompare). La ligne pointillée donne le nombre attendu d'isoformes pour chaque événement (voir Tableau 4.2).

attendu puisque le fichier GTF donné en entrée de FLAIR contient tous les transcrits de références exprimés dans le jeu de données simulé. La différence de prédiction entre RNA-tailor et FLAIR peut être aussi dûe aux alignements d'entrée de minimap2, pour lesquels nous avons observé une sensibilité plus faible par rapport à exonerate. Il est à noter que les prédictions de exonerate s'améliorent à mesure que le taux d'erreur diminue alors que celle de minimap2 ont des résultats constants. Tandis que la différence entre RNA-tailor et FLAIR augmente en faveur de RNA-tailor. Même si nous fournissons des transcrits en pleine longueur exacts (taux d'erreur de 0% dans les lectures), aucune des méthodes n'est capable de récupérer toutes les isoformes. Cela souligne la difficulté de capturer précisément le signal.

Identification des événements de IR et EC artificiels. Pour l'identification des événements artificiels IR et EC, RNA-tailor surpasse systématiquement les deux versions de FLAIR (barres violettes et rouges sur la Fig. 4.11). Cette plus grande sensibilité de RNA-tailor peut être attribuée à sa capacité à extraire, corriger et préserver le signal fourni par exonerate. FLAIR non guidé préserve le signal grâce à une assignation efficace des lectures sur de grandes zones d'alignement et obtient de meilleurs résultats que FLAIR guidé. En effet, la grande taille de ce type d'événements n'endigue pas la non-reconnaissance des événements par l'algorithme de FLAIR guidé. Le nombre d'événements EC ou IR détectés est donc significativement réduit. L'étape `flair-collapse`, qui effondre les isoformes prédites vers les transcrits de référence, est trop forcée, ce qui conduit au rejet des isoformes portant des événements IR ou EC qui ne dépassent pas le seuil de support de lecture par défaut. isONcorrect a un impact minimal sur la reconnaissance des transcrits modifiés car ces événements couvrent de grandes régions de l'ARNm avec peu de variabilité aux

sites d'épissage, protégeant contre la sur-correction. En résumé, l'analyse des résultats de prédiction pour ces deux types d'événements met en évidence la capacité de RNA-tailor, soutenue par la sensibilité de exonerate, à distinguer le signal correct du bruit, une capacité essentielle pour une identification précise et sensible des isoformes.

Identification des événements de ASS artificiels. La reconnaissance des événements ASS est certainement la tâche la plus difficile pour les outils d'identification d'isoformes alternatives, car ces événements concernent des régions très petites (quelques bases) autour des jonctions d'épissage. Les erreurs de séquençage et d'alignement se produisant dans ces zones impactent de manière critique la détection de la bonne jonction d'épissage et, par conséquent, la découverte d'isoformes alternatives. Ainsi, décoder le signal à partir de données bruitées est difficile. Les résultats de la section 4.3.1 ont conclu que le nombre de sites ASS reconnus est similaire entre minimap2 et exonerate. Par conséquent, les données d'entrée pour FLAIR et RNA-tailor sont assez similaires. RNA-tailor identifie autant d'événements ASS que FLAIR non guidé dans chaque expérience (Fig. 4.11, barres bleues). L'application de la correction des lectures avec RNA-tailor supprime environ 20% des événements ASS identifiés. Ce résultat était attendu car préserver les ASS tout en appliquant la correction des lectures est difficile en raison de leur petite taille et de leur impact direct sur la sélection des sites d'épissage. Le résultat le plus notable est l'absence complète de reconnaissance de nouveaux sites ASS avec FLAIR guidé. Cela nous amène à penser que la version guidée de FLAIR n'est pas adaptée à la recherche de nouveaux sites ASS, contrairement à sa version non guidée.

Isoformes prédits partiellement retrouvés (ISM). Pour conclure, notre étude se concentre sur les ISM. Le nombre de ses ISM fournit une bonne approximation des prédictions manquées. La distribution des ISM est notable dans les données brutes des deux aligneurs. En effet, pour celles-ci, l'addition du nombre de FSM et ISM approche ou atteint le nombre maximal de prédiction FSM. Ce signal d'isoformes incomplets, qui peut porter l'événement simulé, est complètement supprimé par l'étape de collapse de FLAIR, tandis qu'il est largement préservé par RNA-tailor. La correction des lectures a tendance à diminuer le nombre de FSM et à augmenter le nombre de ISM. Certaines longues lectures identifiant des transcrits complets sont corrigées, mais les lectures plus courtes issues de la simulation des mêmes isoformes sont conservées et préservent l'événement.

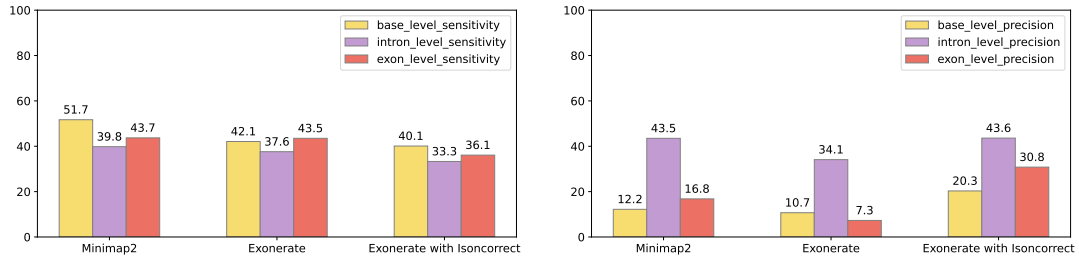
4.4 Évaluation des prédictions d'outils d'identification de transcriptome sur des données réelles

Nous nous intéressons ici à l'analyse de la prédiction d'isoformes à partir du jeu de données SRR15899612 complété par le jeu de données lectures courtes SRR15899613 pour la validation de jonctions d'épissage.

4.4.1 Validation des SJ en sortie d'aligneur : exonerate vs minimap2

Nous commençons par analyser l'efficacité de minimap2 et exonerate sur le jeu de données réel, puis les isoformes prédits par RNA-tailor, FLAIR et Freddie.

Capacité à trouver les bonnes jonctions d'épissage. Nous commençons par une validation croisée des jonctions d'épissage prédites grâce aux courtes lectures. La Figure 4.12 montre comment minimap2 et exonerate capturent les bonnes jonctions d'épissage en supposant que les lectures courtes constituent la référence à laquelle se comparer. Nous avons uniquement testé si une jonction d'épissage prédite (ou intron, ou exon) est supportée par une lecture courte, sans tester les isoformes entiers.



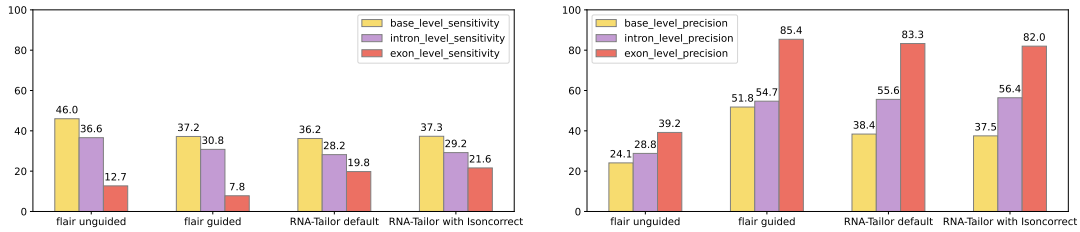
(a) Sensibilité des jonctions d'épissage, introns et exons

(b) Précision des jonctions d'épissage, introns et exons

FIGURE 4.12 – Sensibilité et précision de l'identification des jonctions d'épissage, introns et exons pour minimap2, exonerate avec et sans correction d'erreurs par isONcorrect sur le jeu de données réel comparé aux données de lectures courtes. Une jonction d'épissage identifiée correspond à une jonction d'épissage trouvée dans les alignements *splicé* de minimap2 ou exonerate supportée par au moins une lecture courte. GffCompare est utilisé pour calculer la sensibilité et la précision.

Nous observons que minimap2 et exonerate suivent une tendance similaire en termes de sensibilité globale. La seule différence est que la sensibilité de base de minimap2 est légèrement supérieure à celle de exonerate et que la précision au niveau des exons de exonerate s'améliore beaucoup avec la correction par isONcorrect. Ainsi, les données utilisées par RNA-tailor et FLAIR ont une qualité d'alignement globale similaire selon la validation par courtes lectures.

La Figure 4.13 montre les mêmes statistiques que la Figure 4.12 mais pour comparer FLAIR (guidé et non guidé) et RNA-tailor (avec et sans correction de lecture).



(a) Sensibilité des jonctions d'épissage, introns et exons

(b) Précision des jonctions d'épissage, introns et exons

FIGURE 4.13 – Sensibilité et précision de la récupération des jonctions d'épissage pour FLAIR (guidé et non guidé) et RNA-tailor (avec et sans correction de lecture) sur le jeu de données SRR15899612 Une jonction d'épissage récupérée correspond à une jonction d'épissage trouvée dans les isoformes prédits et supportée par au moins une courte lecture. GffCompare est utilisé pour calculer la sensibilité et la précision.

En termes de sensibilité, comme précédemment, il y a une tendance similaire entre les deux outils. La version non guidée de FLAIR obtient de meilleurs résultats que la version guidée, montrant une meilleure capacité à découvrir de nouveaux événements, mais elle obtient des résultats moins bons que les trois autres outils en termes de précision. RNA-tailor obtient les meilleurs résultats pour la détection des exons en termes de sensibilité et, bien que RNA-tailor n'utilise pas d'annotations en entrée, il obtient des performances similaires à celles de FLAIR guidé. RNA-tailor obtient de meilleurs résultats au niveau des exons et FLAIR guidé obtient de meilleurs résultats en précision au niveau de base.

Capacité à trouver de bons isoformes. Nous nous interrogeons ici sur la viabilité de nos transcrits et définissons un transcrit viable comme étant un transcrit qui possède une ORF. Cependant, les positions des régions UTR n'étant pas connues, nous choisissons de compter le nombre de codons stop dans les exons internes de chaque transcrit prédit. En utilisant la séquence interne des transcrits prédits comme proxy et en évaluant le nombre de codons stop sur chaque phase de cette séquence interne, nous tentons de déterminer si les transcrits ont une phase protéique. La Figure 4.14 montre le ratio des isoformes prédits pour lesquels il y a une phase sans codon STOP dans les exons internes pour RNA-tailor (avec et sans correction de lecture), Freddie et FLAIR (guidé et non guidé). Le calcul pour les 875 transcrits connus a été ajouté comme contrôle. Les transcrits ayant moins de 3 exons ont été filtrés. Bien que le ratio devrait être de 1 pour les transcrits connus, nous observons que certains transcrits contiennent un codon STOP dans les exons internes. Cela provient de transcrits dont le CDS commence après le premier exon ou se termine après le dernier exon.

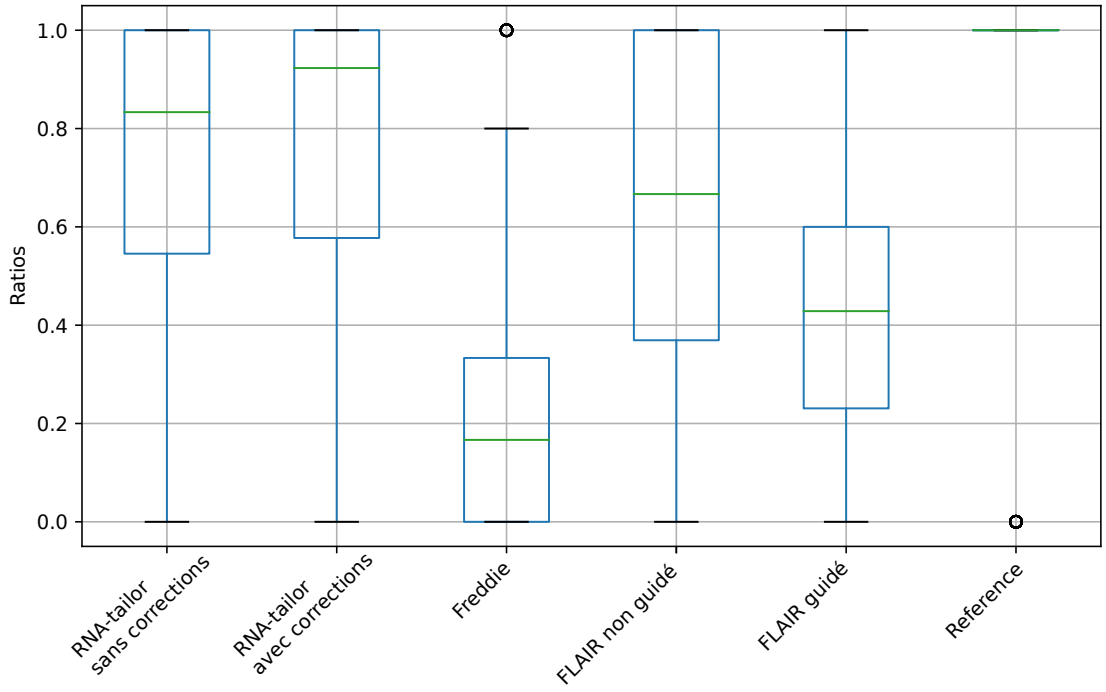


FIGURE 4.14 – Distribution du ratio des isoformes prédits par gène pour lesquels il y a une phase sans codon STOP dans les exons internes pour RNA-tailor (avec et sans correction de lecture), Freddie et FLAIR (guidé et non guidé), et les 875 transcrits connus. Seuls les transcrits avec au moins 3 exons ont été analysés.

Nous observons que la distribution des ratios pour les deux versions de RNA-tailor obtient de meilleurs résultats, avec la médiane la plus proche de 1. Globalement, cela suggère que RNA-tailor a un niveau de précision plus élevé pour prédire précisément toutes les jonctions d'épissage d'un transcrit. De manière intéressante, FLAIR non guidé montre de meilleures performances que FLAIR guidé. Cela peut être lié aux résultats de la Figure 4.13, où FLAIR non guidé a une sensibilité plus élevée que FLAIR guidé dans tous les critères mais une précision moindre. Nous soulignons que l'utilisation d'annotations introduit un biais fort dans la prédiction des isoformes, ce qui entraîne la fusion d'isoformes potentiellement FSM pendant l'étape de correction de FLAIR. En effet, toutes les jonctions d'épissage prédites étant à proximité d'une jonction connue sont corrigées à la jonction

connue, (sous-section 2.3.3). Comme prévu par les comptes nuls de FSM, les résultats pour Freddie mettent en évidence un problème de précision des prédictions des jonctions d'épissage.

Pour terminer l'analyse sur le jeu de données réel, et bien que nous n'ayons pas de vérité terrain pour cela, nous avons décidé de comparer les isoformes prédits avec les annotations connues. Les résultats présentés devraient donc être comparés entre les outils, et non interprétés comme la capacité d'une méthode à trouver les isoformes réellement exprimés. Nous avons utilisé GffCompare pour calculer le nombre de FSM par rapport à la référence connue (875 isoformes). Cela a été fait sur tous les isoformes prédits mais aussi sur un sous-ensemble de ceux-ci, en ne conservant que ceux entièrement supportés par les alignements de courtes lectures (c'est-à-dire les isoformes prédits ayant toutes leurs jonctions d'épissage supportées). Les résultats de cette expérience sont présentés dans le Tableau 4.4.

TABLE 4.4 – Nombre de FSM par rapport à la référence connue et nombre total de transcrits prédits et pour tous les transcrits prédits le sous-ensemble ayant toutes leurs jonctions d'épissage supportées par des courtes lectures. La quatrième colonne donne le ratio de FSM dans le sous-ensemble. Les faux positifs potentiels (FP) sont les transcrits prédits non inclus dans le sous-ensemble. La dernière colonne donne le ratio de transcrits prédits n'ayant pas toutes leurs jonctions d'épissage supportées par des courtes lectures.

	FSM		$\frac{\text{filtrés}}{\text{tous}}$	FP	isoformes prédits		supprimés
	tous	filtrés		potentiels	tous	filtrés	
RNA-tailor	345	333	96.5%	12	3421	2933	14.2%
RNA-tailor avec correction	347	331	95.3%	16	4198	3319	20.9%
FLAIR non guidé	323	294	91.0%	29	5857	2432	58.4%
FLAIR guidé	424	380	89.6%	44	1554	1052	32.3%

Le nombre le plus élevé de FSM récupérés est atteint par FLAIR guidé, tandis que la proportion de FSM dont toutes les jonctions d'épissage sont supportées par des courtes lectures est la plus faible. Les deux versions de RNA-tailor obtiennent des résultats légèrement meilleurs en termes de récupération des FSM que FLAIR non guidé. FLAIR non guidé obtient les pires performances en termes d'isoformes prédits dont toutes les jonctions d'épissage ne sont pas supportées par des courtes lectures. En considérant uniquement les isoformes dont toutes les jonctions d'épissage sont supportées par des courtes lectures, RNA-tailor prédit un nombre plus élevé de FSM, démontrant sa capacité à capturer des structures introniques précises au niveau de base. De plus, le pourcentage d'isoformes filtrés par le support des courtes lectures montre que RNA-tailor, bien qu'il ne soit pas guidé, produit plus d'isoformes prédits supportés dans leur alignement et donc moins susceptibles de produire des faux positifs. Cette expérience met en évidence le problème de l'incomplétude des bases de données. Une méthode qui prend en compte les isoformes connus peut avoir tendance à sélectionner des isoformes moins bien supportés par l'alignement pour se rapprocher d'annotations connues.

4.4.2 Variabilités des prédictions en isoforme FSM.

La résolution de l'inventaire des isoformes présents dans chaque gène est un problème difficile. Néanmoins, on peut s'attendre à ce que les isoformes prédits par différents outils à partir des mêmes lectures ou fichiers d'alignement soient cohérents. Pour vérifier cette hypothèse, on compare la composition des isoformes prédits par RNA-tailor, FLAIR et Isoquant à partir des données réelles et filtrées par les lectures courtes. Freddie n'a pas été gardé car il ne possédait pas assez d'isoformes connus comparés aux autres méthodes. Cependant, on a choisi d'ajouter Isoquant pour ajouter un exemple de set de prédictions et souligner la variabilité des isoformes prédits à partir des mêmes données.

TABLE 4.5 – Table récapitulatif du nombre d'isoformes connus identifiés par Isoquant, FLAIR et RNA-tailor après validation croisée par SRR15899613. Le nombre cumulé correspond au nombre d'isoformes uniques total retrouvés par l'ensemble des méthodes.

	RNA-Tailor	Isoquant	Flair	Cumulé
FSM	347	247	323	438
autres (y compris ISM)	4198	590	5837	

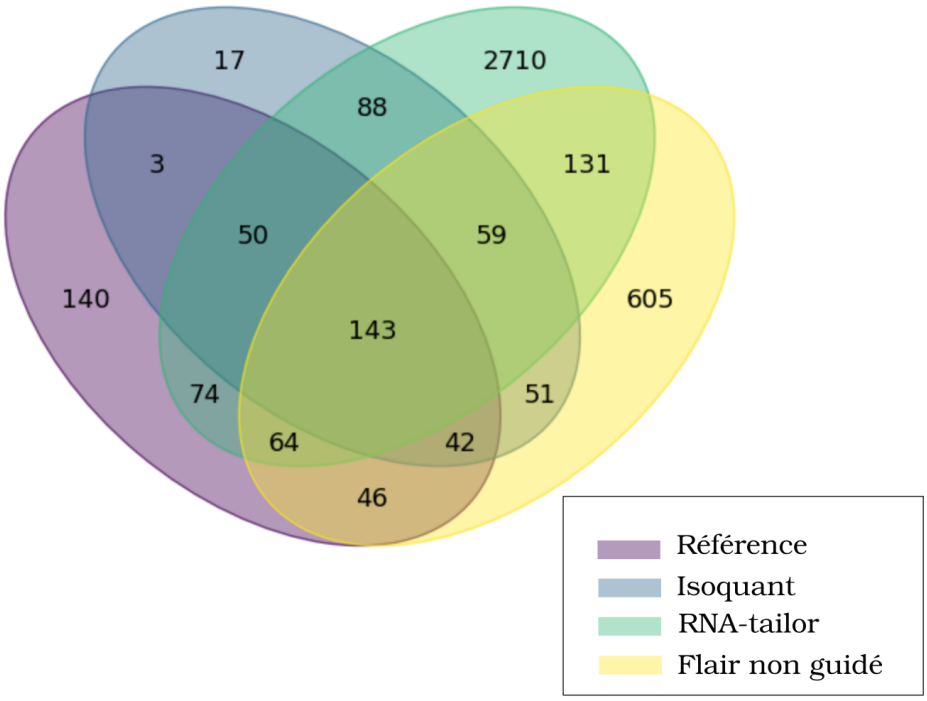


FIGURE 4.15 – Comparaison des compositions en isoformes des prédictions de trois outils : RNA-tailor, FLAIR non guidé et Isoquant. Tous les résultats de prédictions ont été réalisés à partir du jeu de données SRR15899612. La référence des 875 isoformes et les résultats de prédictions ont été filtrés par le jeux de donnée SRR15899613 comme décrit en figure 4.12, afin de ne garder que les isoformes supportés par des lectures courtes. Après validation croisée, on conserve 562 isoformes de référence.

On observe que chaque outil prédit des isoformes uniques dont la structure correspond à un isoforme connu soutenu par les lectures courtes. Ils représentent au total 123 isoformes (Isoquant (3), RNA-tailor (74) et FLAIR (46)). Le total cumulé des isoformes connus est supérieur au nombre de prédictions réalisés par chaque outil (26% de plus que RNA-tailor, 35% que FLAIR et 77% pour Isoquant). Ici, sur les données réelles, tout indique que l'existence de ses isoformes est fiable. Cependant, les différences de composition entre outils sont frappantes. RNA-tailor possède 21% de prédictions propres contre 14% pour FLAIR et 1% pour Isoquant. La prédiction des isoformes est sujette à une grande variabilité selon les outils utilisés, ce qui sous-entend qu'un consensus en matière de méthode de prédiction n'a pas encore été trouvé.

4.5 Approche exploratoire pour les gènes étudiés.

L'objectif de cette partie est d'étudier l'apport de chaque étape de raffinement des résultats sur la qualité des prédictions des isoformes.

4.5.1 Effet des méthodes de raffinage des prédictions de RNA-tailor

Pour ce faire, nous allons observer le nombre de points de jonctions (JPs) prédits après chacune des étapes de raffinement des alignements. Le nombre de points de jonctions est un bon marqueur de la précision des résultats de prédictions parce qu'elles permettent d'apprécier la précision du choix des sites d'épissage dans leur globalité.

Le tableau 4.6 reflète l'effet de chaque méthode de correction sur les résultats de prédiction des transcrits isoformes. Dans ce tableau, on suit l'évolution des points de jonctions

JPs dans la référence : 8864	JPs prédits	JPs prédits communs	JPs prédits spécifiques
en sortie de exonerate	18445	7492	10953
après préfiltrage	13832	7492	6340
après réalignement	14145	7470	6675
après lissage	11653	7459	4194

TABLE 4.6 – Table de l'évolution de la composition en point de jonction (JPs), i.e. site d'épissage donneur ou accepteur, par rapport à l'ensemble des JPs uniques contenus dans la référence lectures courtes SRR15899613 (8864). Ici, les résultats proviennent de l'analyse par RNA-tailor des données SRR15899612 pour les 280 gènes sélectionnés. Les JPs prédits correspondent aux nombres totaux de JPs uniques observés dans les isoformes solides prédits après étape de RNA-tailor. Les JPs prédits communs sont la partie des JPs uniques retrouvés dans les JPS de la référence. A l'inverse, les JPs prédits spécifiques sont les JPs prédits non retrouvés dans la référence.

communs avec la référence et des autres points de jonctions prédits (qu'on pourrait qualifier de faux positifs). En sortie de exonerate, le nombre de JP prédits est maximal. Le pré-filtrage a pour rôle d'éliminer les JP trop peu supportés. Il induit la suppression de 4619 JP soit 41% des JPs tout en préservant l'entièreté du signal supporté par le séquençage Illumina, puisqu'aucun JP prédit commun n'est supprimé. Le réalignement, quant à lui, réaligne des portions de reads contenus dans des blocs faiblement soutenus grâce à un alignement local. Cet alignement est très précis et peut être sensible aux erreurs de séquençage à son tour. Ainsi il a pour effet de bord d'après l'analyse des JPs d'augmenter le nombre de JPs total prédit. Il réintroduit du bruit de séquençage. Un exemple de cette réintroduction de variabilité dans les JPs est montré dans la figure 4.16c. Ce bruit réintroduit peu ensuite être corrigé par le lissage. Le lissage réduit de 38% supplémentaire le nombre de JP uniques par rapport au réalignement. Ainsi, les méthodes de correction

consécutives ont un fort impact sur la réduction de l'hétérogénéité des JP contenus dans les prédictions en sortie de exonerate.

4.5.2 Deux cas particuliers

Illustration du préfiltrage. Le préfiltrage n'a pas d'action directe sur la reconnaissance précise des sites d'épissage mais une action sur la reconnaissance et l'élimination de de lectures contaminantes. Cette dernière peut engendrer des conséquences sur la suite de l'analyse. Par exemple, dans la table 4.7, l'élimination de 5 lectures par la méthode du filtrage engendre la baisse du seuil de reconnaissance des isoformes solides d'un support demandé de 3 lectures à 2 lectures. Ce changement de seuil sélectionne 9 isoformes solides supplémentaire, dont 1 FSM.

	Nombre de structures introniques prédites totales	Nombre de structures introniques prédites solides	FSM
avant préfiltrage	57	12	2
après préfiltrage	52	21	3

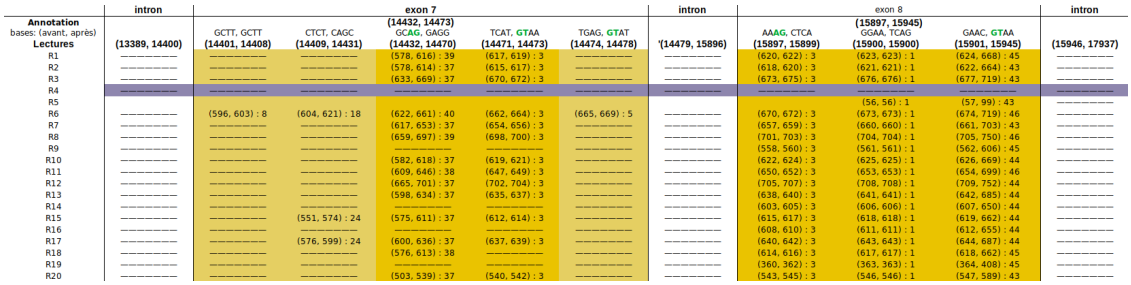
TABLE 4.7 – Effet du pré-filtrage sur la composition en isoformes pour le gène ENSG00000110031. Le pré-filtrage induit la suppression de lectures dites fragiles, c'est-à-dire ayant une structure d'épissage possédant plus de 25% de points de jonctions fragiles.

Illustration du réalignement. Comme aperçu dans le tableau 4.6, le nombre de points de jonction uniques présents dans le jeu de données est affecté par chaque étape de raffinement des données. La figure 4.16 met en évidence les variations de structures introniques induites par chacune de ces étapes. Le réalignement joue un rôle important dans la zone sélectionnée. La partie (14401,14431) de l'exon 7 est soutenue par deux blocs fragiles. Les fragments de lecture contenus dans ces blocs sont réalignés sur les positions génomiques (13363, 13388) (la totalité de la région génomique n'est pas visible sur la figure 4.16). Ces réalignements permettent de replacer précisément la borne solide de l'exon prédit 7. Entre les blocs solides des exons 7 et 8, trois réalignements ont lieu. La portion (596,621) de R6 et les portions (551, 574) et (576, 599) de R14 et R17 alignées sur les blocs fragiles ((14401, 14408) et (14409, 14431)) ont été réalignées dans l'exon 6 sur le bloc (13363, 13388). Pour finir, une portion de la lecture R4 plus distante en amont est réalignée sur l'exon 8. Son réalignement provoque la formation d'un bloc supplémentaire en (15946, 15947). Le réalignement de cette sous portion de lecture a donc mené à la création d'un point de jonction supplémentaire dans nos données. L'alignement de la lecture R4 est montré en figure 4.17. L'alignement orignal par exonerate avec le modèle *est2genome* est montré dans la sous-figure 4.17a. On observe que la portion identifiée pour le réalignement, encadrée en rouge, contient beaucoup d'erreurs d'alignements et est soutenue par un site d'épissage faible 'ga'. Le réalignement puis le lissage permet de replacer correctement la sous-portion qui est supportée par un site donneur 'gt' en amont ('gt' en surligné en rouge sur la figure 4.17a et en aval du réalignement grâce au lissage en 15945 en figure 4.16c en vert pour la lecture R4).

Le réalignement permet donc de corriger des alignements de lectures mais aussi de garder de la diversité dans les données. Ce second point est permis par le fait de se placer à l'échelle du gène pour l'utilisateur.

4.6 Réflexions sur l'amélioration des résultats par post-traitement des isoformes prédits

RNA-tailor possède une philosophie différente des autres outils. Il se place à l'échelle du gène et son export XLSX a été développé pour faciliter les approches exploratoires.



Annotation bases: (avant, après) Lectures	intron (13389, 14400)	exon 7 (14432, 14473) GCTT, GCTT (14401, 14408)	CTCT, CAGC (14409, 14431)	GCAG, GAGG (14432, 14470)	TCAT, GTAA (14471, 14473)	TGAG, GTAT (14474, 14478)	intron '(14479, 15896)	exon 8 (15897, 15945) AAAG, CTCA (15897, 15899)	GGAA, TCAG (15900, 15900)	GAAC, GTAA (15901, 15945)	intron (15946, 17937)
R1	————	————	————	(578, 616) : 39	(617, 619) : 3	————	————	(620, 622) : 3	(623, 623) : 1	(624, 668) : 45	————
R2	————	————	————	(578, 614) : 37	(615, 617) : 3	————	————	(618, 620) : 3	(621, 621) : 1	(622, 664) : 43	————
R3	————	————	————	(633, 669) : 37	(670, 672) : 3	————	————	(673, 675) : 3	(676, 676) : 1	(677, 719) : 43	————
R4	————	————	————	————	————	————	————	————	————	————	————
R5									(56, 56) : 1	(57, 99) : 43	————
R6	————	(596, 603) : 8	(604, 621) : 18	(622, 661) : 40	(662, 664) : 3	(665, 669) : 5	————	(670, 672) : 3	(673, 673) : 1	(674, 719) : 46	————
R7	————	————	————	(617, 653) : 37	(654, 656) : 3	————	————	(657, 659) : 3	(660, 660) : 1	(661, 703) : 43	————
R8	————	————	————	(659, 697) : 39	(698, 700) : 3	————	————	(701, 703) : 3	(704, 704) : 1	(705, 750) : 46	————
R9	————	————	————	————	————	————	————	(558, 560) : 3	(561, 561) : 1	(562, 606) : 45	————
R10	————	————	————	(582, 618) : 37	(619, 621) : 3	————	————	(622, 624) : 3	(625, 625) : 1	(626, 669) : 44	————
R11	————	————	————	(609, 646) : 38	(647, 649) : 3	————	————	(650, 652) : 3	(653, 653) : 1	(654, 699) : 46	————
R12	————	————	————	(665, 701) : 37	(702, 704) : 3	————	————	(705, 707) : 3	(708, 708) : 1	(709, 752) : 44	————
R13	————	————	————	(598, 634) : 37	(635, 637) : 3	————	————	(638, 640) : 3	(641, 641) : 1	(642, 685) : 44	————
R14	————	————	————	————	————	————	————	(603, 605) : 3	(606, 606) : 1	(607, 650) : 44	————
R15	————	————	(551, 574) : 24	(575, 611) : 37	(612, 614) : 3	————	————	(615, 617) : 3	(618, 618) : 1	(619, 662) : 44	————
R16	————	————	————	————	————	————	————	(608, 610) : 3	(611, 611) : 1	(612, 655) : 44	————
R17	————	————	(576, 599) : 24	(600, 636) : 37	(637, 639) : 3	————	————	(640, 642) : 3	(643, 643) : 1	(644, 687) : 44	————
R18	————	————	————	(576, 613) : 38	————	————	————	(614, 616) : 3	(617, 617) : 1	(618, 662) : 45	————
R19	————	————	————	————	————	————	————	(360, 362) : 3	(363, 363) : 1	(364, 408) : 45	————
R20	————	————	————	(503, 539) : 37	(540, 542) : 3	————	————	(543, 545) : 3	(546, 546) : 1	(547, 589) : 43	————

(a) Structure de la zone d'alignement du gène avant réalignement.

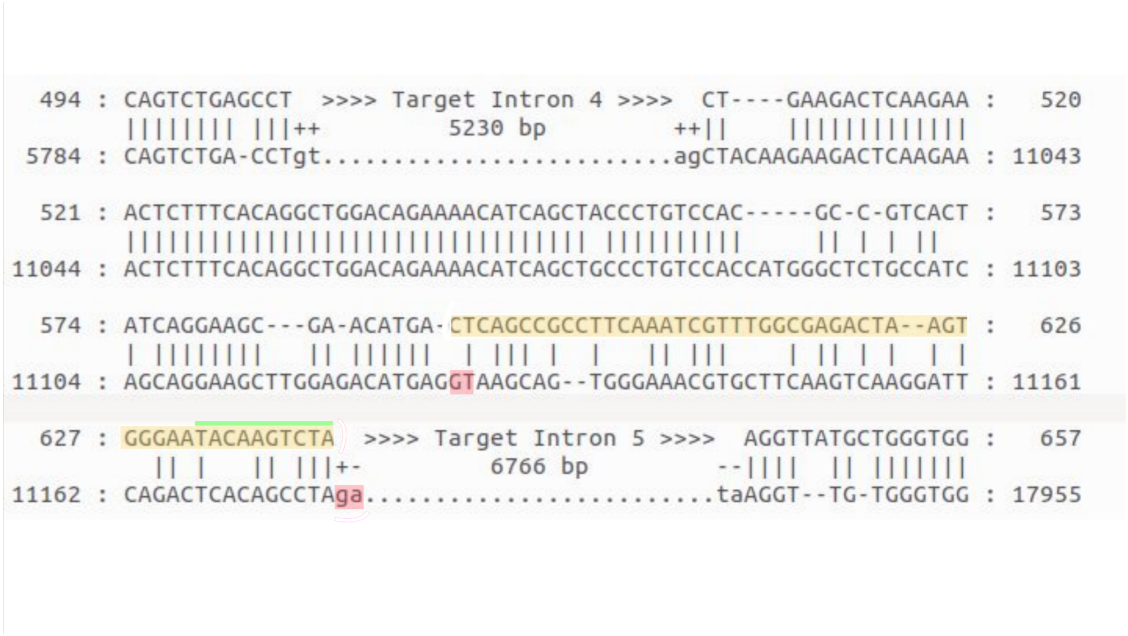
Annotation bases: (avant, après) Lectures	intron (13389, 14431)	exon 7 (14432, 14473) GCAG, GAGG (14432, 14470)	TCAT, GTAA (14471, 14473)	TGAG, GTAT (14474, 14478)	intron '(14479, 15896)	exon 8 (15897, 15945) AAAG, CTCA (15897, 15899)	GGAA, TCAG (15900, 15900)	GAAC, GTAA (15901, 15945)	AAAG, AACA (15946, 15947)	intron (15948, 17937)
R1	————	(578, 616) : 39	(617, 619) : 3	————	————	(620, 622) : 3	(623, 623) : 1	(624, 668) : 45	————	————
R2	————	(578, 614) : 37	(615, 617) : 3	————	————	(618, 620) : 3	(621, 621) : 1	(622, 664) : 43	————	————
R3	————	(633, 669) : 37	(670, 672) : 3	————	————	(673, 675) : 3	(676, 676) : 1	(677, 719) : 43	————	————
R4	————	————	————	————	————	————	(591, 592) : 2	(593, 638) : 46	(639, 641) : 3	————
R5							(56, 56) : 1	(57, 99) : 43	————	————
R6	————	(622, 661) : 40	(662, 664) : 3	(665, 669) : 5	————	(670, 672) : 3	(673, 673) : 1	(674, 719) : 46	————	————
R7	————	(617, 653) : 37	(654, 656) : 3	————	————	(657, 659) : 3	(660, 660) : 1	(661, 703) : 43	————	————
R8	————	(659, 697) : 39	(698, 700) : 3	————	————	(701, 703) : 3	(704, 704) : 1	(705, 750) : 46	————	————
R9	————	————	————	————	————	(558, 560) : 3	(561, 561) : 1	(562, 606) : 45	————	————
R10	————	(582, 618) : 37	(619, 621) : 3	————	————	(622, 624) : 3	(625, 625) : 1	(626, 669) : 44	————	————
R11	————	(609, 646) : 38	(647, 649) : 3	————	————	(650, 652) : 3	(653, 653) : 1	(654, 699) : 46	————	————
R12	————	(665, 701) : 37	(702, 704) : 3	————	————	(705, 707) : 3	(708, 708) : 1	(709, 752) : 44	————	————
R13	————	(598, 634) : 37	(635, 637) : 3	————	————	(638, 640) : 3	(641, 641) : 1	(642, 685) : 44	————	————
R14	————	————	————	————	————	(603, 605) : 3	(606, 606) : 1	(607, 650) : 44	————	————
R15	————	(575, 611) : 37	(612, 614) : 3	————	————	(615, 617) : 3	(618, 618) : 1	(619, 662) : 44	————	————
R16	————	(566, 600) : 35	(601, 603) : 3	(604, 607) : 4	————	(608, 610) : 3	(611, 611) : 1	(612, 655) : 44	————	————
R17	————	(600, 636) : 37	(637, 639) : 3	————	————	(640, 642) : 3	(643, 643) : 1	(644, 687) : 44	————	————
R18	————	(576, 613) : 38	————	————	————	(614, 616) : 3	(617, 617) : 1	(618, 662) : 45	————	————
R19	————	————	————	————	————	(360, 362) : 3	(363, 363) : 1	(364, 408) : 45	————	————
R20	————	(503, 539) : 37	(540, 542) : 3	————	————	(543, 545) : 3	(546, 546) : 1	(547, 589) : 43	————	————

(b) Structure de la zone d'alignement du gène après réalignement.

Annotation bases: (avant, après) Lectures	intron (13389, 14431)	exon 7 (14432, 14473) GCAG, GAGG (14432, 14470)	TCAT, GTAA (14471, 14473)	TGAG, GTAT (14474, 14478)	intron '(14479, 15896)	exon 8 (15897, 15945) AAAG, CTCA (15897, 15899)	GGAA, TCAG (15900, 15900)	GAAC, GTAA (15901, 15945)	intron (15946, 17937)
R1	————	(578, 616) : 39	(617, 619) : 3	————	————	(620, 622) : 3	(623, 623) : 1	(624, 668) : 45	————
R2	————	(578, 614) : 37	(615, 617) : 3	————	————	(618, 620) : 3	(621, 621) : 1	(622, 664) : 43	————
R3	————	(633, 669) : 37	(670, 672) : 3	————	————	(673, 675) : 3	(676, 676) : 1	(677, 719) : 43	————
R4	————	————	————	————	————	————	(591, 592) : 2	(593, 641) : 49	————
R5							(56, 56) : 1	(57, 99) : 43	————
R6	————	(622, 661) : 40	(662, 664) : 3	(665, 669) : 5	————	(670, 672) : 3	(673, 673) : 1	(674, 719) : 46	————
R7	————	(617, 653) : 37	(654, 656) : 3	————	————	(657, 659) : 3	(660, 660) : 1	(661, 703) : 43	————
R8	————	(659, 697) : 39	(698, 700) : 3	————	————	(701, 703) : 3	(704, 704) : 1	(705, 750) : 46	————
R9	————	————	————	————	————	(558, 560) : 3	(561, 561) : 1	(562, 606) : 45	————
R10	————	(582, 618) : 37	(619, 621) : 3	————	————	(622, 624) : 3	(625, 625) : 1	(626, 669) : 44	————
R11	————	(609, 646) : 38	(647, 649) : 3	————	————	(650, 652) : 3	(653, 653) : 1	(654, 699) : 46	————
R12	————	(665, 701) : 37	(702, 704) : 3	————	————	(705, 707) : 3	(708, 708) : 1	(709, 752) : 44	————
R13	————	(598, 634) : 37	(635, 637) : 3	————	————	(638, 640) : 3	(641, 641) : 1	(642, 685) : 44	————
R14	————	————	————	————	————	(603, 605) : 3	(606, 606) : 1	(607, 650) : 44	————
R15	————	(575, 611) : 37	(612, 614) : 3	————	————	(615, 617) : 3	(618, 618) : 1	(619, 662) : 44	————
R16	————	(566, 600) : 35	(601, 603) : 3	(604, 607) : 4	————	(608, 610) : 3	(611, 611) : 1	(612, 655) : 44	————
R17	————	(600, 636) : 37	(637, 639) : 3	————	————	(640, 642) : 3	(643, 643) : 1	(644, 687) : 44	————
R18	————	(576, 613) : 38	————	————	————	(614, 616) : 3	(617, 617) : 1	(618, 662) : 45	————
R19	————	————	————	————	————	(360, 362) : 3	(363, 363) : 1	(364, 408) : 45	————
R20	————	(503, 539) : 37	(540, 542) : 3	————	————	(543, 545) : 3	(546, 546) : 1	(547, 589) : 43	————

(c) Structure de la zone d'alignement du gène après lissage.

FIGURE 4.16 – Capture d'écran des fichiers XLSX en sortie de RNA-tailor pour l'analyse du gène ENSMUSG00000000827 à partir des données produites par séquençage MinION (PRJEB25574) durant le projet ASTER. Cet exemple permet de suivre le processus de correction des alignements d'exonerate. Chaque figure est un zoom sur une zone d'intérêt dans laquelle se produisent des réalignements et du lissage. (a) structures introniques en sortie des alignements par exonerate, les lectures présentes sont celles qui ont passé l'étape de pré-filtrage. La zone violette est la zone de réalignement de la lecture R4, (b) le réalignement également affecte les bornes des exons 7 et 8. Ici, le réalignement de R4 sur l'exon 8 (bloc d'alignement en vert et rouge) a créé un bloc supplémentaire (15946, 15947), qui n'est pas soutenu par un site d'épissage. (c) cette structure est raffinée par le lissage des bordures. Pour information, la structure connue de ce gène donne 3 introns sur cette région aux positions (13389, 14431), (14479, 15896), (15946, 17937) qui sont retrouvés par les blocs solides.



(a) Résultat de l'alignement par exonerate (mode est2genome) de la lecture R4 sur la séquence du gène ENSMUS000000827 restreint à la portion 5784 à 17955 du gène.

(b) Résultat du réalignement par exonerate (mode affine :local) de la sous partie (0, 40) de la lecture R4 sur la séquence du gène ENSMUSG00000000827 restreint à la portion 4774 à 4812 du gène.

FIGURE 4.17 – Alignements exonerate en sortie de RNA-tailor pour la lecture R4. La figure (a) représente la portion de R4 dont l'alignement a été identifié pour réalignement (surligné en jaune) et la figure (b) représente le nouvel alignement validé. Les dix bases de la lecture qui sont marquées en vert dans la figure (a) ne sont pas réalignées par le réalignement local en figure b). Cependant, notre méthode accole les bases restantes dans les blocs contigus (blocs de 5+3 base visible en figure 4.16b tout à droite.

De part cette approche à plus basse échelle, RNA-tailor permet une approche exploratoire via la prédiction d'un grand nombre de structure d'isoformes (les solides et les fragiles). Dans son pipeline, une ligne claire sous la forme d'un seuil de support de lecture est défini pour différencier les isoformes prédits solides et fragiles. Cependant, cette dichotomie a ses limites. De part la nature la dégradation des séquençage 3ème génération il se peut que seules certaines lectures possèdent la pleine longueur. Ainsi, des isoformes « vrais » peuvent être peu soutenus et qualifiés de fragiles. On le constate lorsqu'on analyse des jeux de données simulés pour lesquels on connaît la vérité. Par exemple, si on reprend les données réelles du jeu de données SRR15899612, et qu'on analyse les isoformes fragiles de RNA-tailor soutenus par des lectures courtes, on trouve 44 FSM supplémentaires. Parmi eux 19 FSM sont spécifiques à FLAIR, 1 FSM est spécifique à Isoquant et 24 FSM sont spécifiques à RNA-tailor. Ainsi RNA-tailor a classé en tant que fragiles des isoformes identifiés comme de confiance par d'autres méthodes. En perspective, il reste du signal à détecter dans les isoformes fragiles. Le module complémentaire à l'analyse de RNA-tailor propose d'exploiter ces isoformes fragiles via différentes stratégies, dont certaines peuvent être combinées, pour analyser différemment les prédictions solides et fragiles (voir description des stratégies Section 3.6). En particulier trois pistes ont été envisagées : corriger encore plus les bornes des introns prédits via la recherche de signaux d'épissage sur la séquence du gène ; s'appuyer sur les ORF prédites pour regrouper des isoformes ; et le traitement de la problématique liée à l'inclusion de certains isoformes les uns dans les autres, problème dû en partie au séquençage incomplet de certains transcrits. On termine par une réflexion sur le nombre de lectures support pour qualifier un isoforme solide.

4.6.1 Correction des sites d'épissage

Cette méthode de correction des sites d'épissage est inspirée du lissage des bordures. Elle propose une version plus « brute » puisque les bornes sont corrigées de manière autoritaire vers le site canonique le plus proche (dans une fenêtre de deux nucléotides). Cela a l'inconvénient d'effacer complètement les sites alternatifs faibles mais permet des corrections dans le cas d'erreur d'alignement. Dans le tableau 4.8, la correction vers les sites canoniques permet de conserver l'ensemble des FSM prédits pour PBSIM_0. Pour PBSIM_10, la correction permet de racheter 13 structures introniques classés FP vers les FSM, ce qui permet de réduire le nombre de FP. En revanche, il faut rester vigilant quand à l'utilisation de cette technique car elle privilégie uniquement les sites GT/AG et les sites alternatifs faible soutenant des FSM serait supprimés.

		sortie de RNA-tailor	correction GT/AG
PBSIM_0	FSM	862	862
PBSIM_10	FSM	731	744
	FP	866	967

TABLE 4.8 – Variation du nombre de FSM selon l'application de la correction vers les sites d'épissage canoniques.

4.6.2 Regroupement par ORF prédite

Le regroupement par ORF des isoformes (voir Section 3.6.2) permet de conserver l'ensemble de l'information prédite de composition de l'expérience en protéine. Son but est de réduire la part de FP en regroupant les isoformes doublons en terme d'ORF potentiel. Les résultats de l'expérience du regroupement par ORF est visible en table 4.9. Sur PBSIM_0 le groupement par ORF ne conserve pas l'entièreté des isoformes prédits par la méthode. Il existe donc 11 isoformes de référence qui possède un doublon d'ORF chez un autre isoforme

de référence. Ainsi plusieurs isoformes codent pour la même protéine. Sur les données de PBSIM_10, ce regroupement permet de réduire

		sortie de RNA-tailor	regroupement par ORF
PBSIM_0	FSM	862	851
PBSIM_10	FSM	731	739
	FP	866	934

TABLE 4.9 – Variation du nombre de FSM selon l'application du regroupement par ORF.

4.6.3 La problématique de l'inclusion des isoformes prédits

La dégradation des ARNm en entrée du séquenceur a une forte répercussion sur les qualités du séquençage en réduisant le nombre de transcripts avec un séquençage pleine longueur. Notre méthode d'inclusion (Section 3.6.3), mais aussi une option activable de FLAIR (Section 2.3.3) intègre ce constat de manière similaire pour sélectionner les structures susceptibles d'être pleine longueur et de filtrer les autres.

Nous proposons ici d'analyser le résultat de plusieurs stratégies mises en place pour décider qu'un isoforme est inclus ou non dans un autre. Pour étudier l'impact des différentes stratégies, nous avons fait le choix d'analyser le jeu de données simulé sans erreur PBSIM_0 pour servir de référence et PBSIM_10 pour quantifier l'impact.

On commence par l'inclusion simple (telle que décrite Section 3.6.3). L'évolution du nombre de FSM parmi les isoformes solides ainsi que le nombre de faux positifs (isoformes solides n'étant pas dans la base de vérité) est donné dans le tableau 4.10, colonne « inclusion ». Ces premiers résultats montrent un double effet de réduction à la fois sur les

		sortie de RNA-tailor	inclusion	inclusion + statut UTR	inclusion + statut UTR + ORF check
PBSIM_0	FSM	862	752	791	797
PBSIM_10	FSM	731	650	676	678
	FP	866	231	637	638

TABLE 4.10 – Variation du nombre de FSM selon l'application de différentes méthodes de filtrage des isoformes développées dans le module complémentaire de RNA-tailor. La colonne « sortie de RNA-tailor » correspond aux résultats des isoformes solides de RNA-tailor. Chaque signe « + » correspond à l'accumulation d'une stratégie. Les résultats ont été produits sur des lectures parfaites PBSIM_0 (ligne 1) qui ne possèdent que 3 FP, qui ne sont pas suivis ici. Cette ligne permet de représenter l'effet des méthodes dans un contexte idéal. Les lignes 2 et 3 proviennent des résultats de prédictions pour les lectures simulées PBSIM_10.

FSM et les FP. L'inclusion seule a un impact important sur le nombre de FSM et surtout sur le nombre de FP, ceux-ci diminuent respectivement de 11% et de 74%. La réduction importante sur les FP montre que l'hypothèse d'inclusion des isoformes pour éliminer les lectures dégradées est efficace. Un exemple du processus d'inclusion est illustré dans la figure 4.18. Cependant son utilisation induit une limite en terme de sensibilité (752 FSM contre 862 à l'origine pour PBSIM_0, soit 12,8% de perte) qui correspond l'inclusion des isoformes connus les uns dans les autres.

La diminution de la sensibilité (moins de FSM) est problématique. Il faut essayer de protéger des isoformes connus de l'inclusion (on rappelle que l'inclusion appliquée aux données annotées a également pour effet de diminuer le nombre de structures introniques, voir tableau 4.1). On propose alors de contraindre la définition d'isoformes inclus les uns

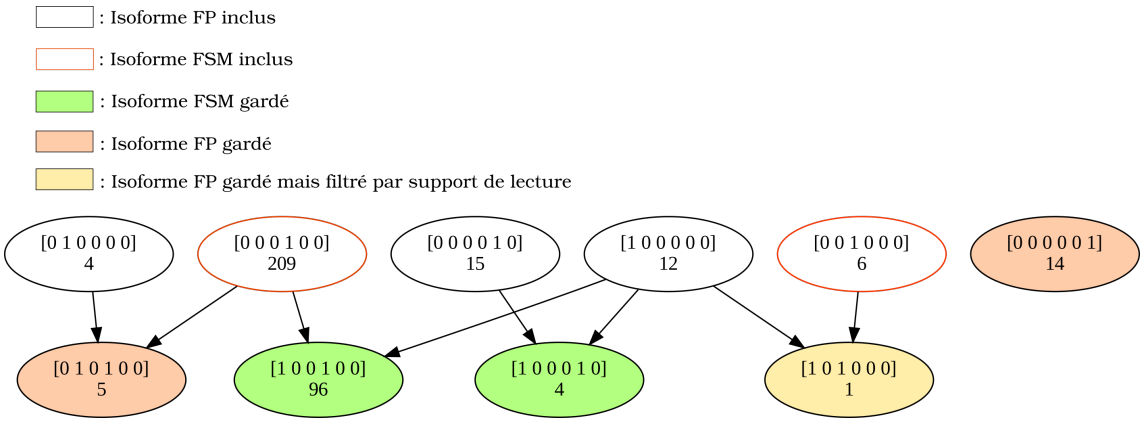
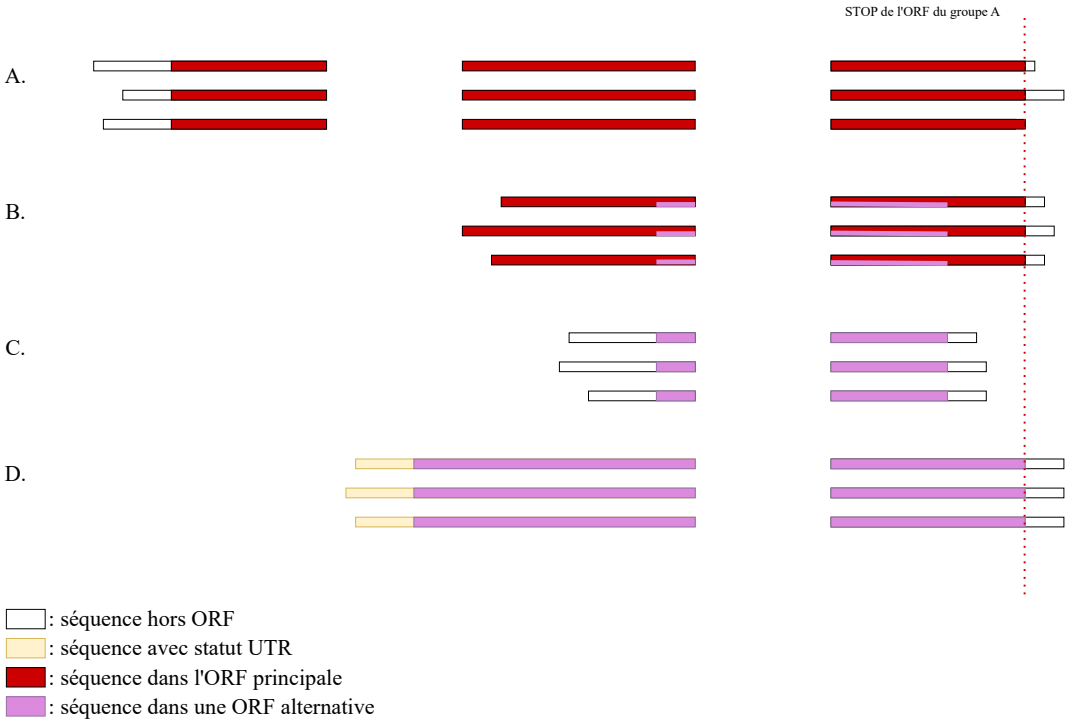


FIGURE 4.18 – Exemple d'inclusion des transcripts prédits par RNA-tailor pour le gène ENSG00000168876, à partir des données SRR15899612 avec isONcorrect. RNA-tailor prédit 10 structures introniques (9 solides + 1 fragile, dont 4 FSM solides). Le processus d'inclusion permet de réduire le nombre de structure de 10 à 5, puis le filtre de support de lecture élimine une structure fragile. Durant le processus d'inclusion, 2 FSM sont inclus dans des FP et donc perdus. Cet exemple met en évidence les problématiques de sélection par inclusion.

dans les autres. L'idée est de dire qu'un isoforme ne peut être inclus dans un autre que si un isoforme possède son premier/dernier bloc fragile alors il ne peut être inclus que dans des isoformes possédant ce même bloc en tant que premier/dernier bloc. D'une certaine manière, cela revient à reconnaître l'incertitude de l'alignement des débuts et fins de lectures qui peuvent représenter des séquences non-transcrites. Ce pourquoi nous qualifierons d'UTR l'ensemble des blocs en amont du premier bloc solide et en aval du dernier bloc solide de chacune des lectures. Cet ajout permet de réduire la perte de FSM dans PBSIM_0 et de conserver 39 FSM supplémentaires. Pour fixer les idées, la figure 4.19 illustre l'effet des différentes stratégies sur la prédiction d'isoformes.

Pour aller plus loin, et dans le même esprit, on peut ajouter à cette contrainte celle de la reconnaissance des ORF (voir Section 3.6.4). On va alors bloquer l'inclusion d'un isoforme A dans un isoforme B si A partage son codon START ou STOP avec B. Cela permet de conserver les isoformes qui ont un ORF alternatif mais pour lequel RNA-tailor n'a pas détecté de bloc UTR. C'est l'exemple de l'isoforme « D » dans la figure 4.19. Ce second ajout permet de réduire encore un peu la perte de FSM dans PBSIM_0 et de conserver 6 FSM supplémentaires dans la table 4.10.

Effet de ces méthodes sur le nombre de prédictions totales et de FSM. Les méthodes de regroupement sur l'ensemble des isoformes font l'effet d'un compromis qui peut être plus ou moins adapté au gène d'intérêt. En effet, leur objectif premier est de réduire le nombre d'isoformes produits pour améliorer la précision de RNA-tailor sans compromettre la sensibilité. L'inclusion est une méthode très efficace mais qui peut masquer des isoformes connus (voir la table 4.10). Elle est préconisée lorsque tous les isoformes connus d'un gène ne sont pas inclus les uns dans les autres. L'ajout de la détection d'UTR et la reconnaissance d'ORF permet de sauver quelques isoformes FSM mais repêche également beaucoup de FP. Le problème de la différenciation des structures FSM et FP est compliqué, il n'existe pas une solution simple et efficace pour traiter ce problème.



(a) Vue schématique de 12 lectures.

Type de regroupement	
Structure intronique	A. B. C. D.
Regroupement par ORF	A. B. C. D.
Inclusion	A. B. C. D.
Inclusion avec statut UTR	A. B. C. D.
Inclusion avec statut UTR et ORF	A. B. C. D.

(b) Regroupements possibles suivant différents critères.

FIGURE 4.19 – Illustration de l'effet des différentes méthodes d'inclusion sur la composition des isoformes prédits en sortie. (a) 12 lectures sont représentées, une par ligne. Les lectures sont regroupées par ensembles d'isoformes putatifs. Les lectures du groupe B, possèdent un STOP commun avec les lectures du groupe A mais n'ont pas de START dans l'ORF de ce groupe. Les lectures du groupe C possèdent une ORF alternative qui diffère pour son START et STOP. Les lectures du groupe D ont un START alternatif par rapport au groupe A et partage son codon STOP. (b) les isoformes prédits en fonction des différentes stratégies. En sortie de RNA-tailor la prédiction est de 2 isoformes l'isoforme A avec 2 introns soutenus par 3 lectures, et l'isoforme (B,C,D) avec 1 intron soutenu par 9 lectures. En appliquant un regroupement par ORF on obtient 3 isoformes (A,(BC),D). En appliquant une inclusion simple, on prédit 1 seul isoforme soutenu par 12 lectures, alors qu'en appliquant une inclusion prenant en compte le statut UTR on obtient deux isoformes (A,B,C) et D car il est protégé par la reconnaissance soit du statut UTR des blocs dans RNA-tailor, soit par la reconnaissance d'un ORF correspondant à un start alternatif.

Conclusion

Les travaux réalisés dans cette étude ont permis de développer et d'évaluer le pipeline RNA-tailor pour l'identification des isoformes épissés alternativement, en le comparant à d'autres outils de référence comme FLAIR, Freddie ou IsoQuant. Les résultats montrent que RNA-tailor présente des avantages en termes de sensibilité et de précision. RNA-tailor a démontré une bonne capacité à identifier les isoformes connus (FSM) par rapport aux autres outils testés. Cette performance est particulièrement visible dans la reconnaissance des événements d'épissage alternatif, où RNA-tailor surpasse souvent FLAIR, notamment pour les événements de rétention d'intron (IR) et d'exon cassette (EC). L'évaluation des performances sur des jeux de données simulés avec différents taux d'erreur de séquençage a montré que RNA-tailor et son aligneur exonerate maintiennent une sensibilité élevée même à des taux d'erreur élevés (10%). La correction des lectures avec isONcorrect a un impact minimal sur la reconnaissance des transcrits modifiés, ce qui montre la robustesse du pipeline face aux erreurs de séquençage. Les analyses ont montré que l'utilisation de minimap2 pour la sélection des lectures, en comparaison avec megablast, permet de capturer un plus grand nombre de lectures pertinentes, augmentant ainsi le support des isoformes connus sans introduire de nouvelles structures erronées. Les méthodes de filtrage développées dans RNA-tailor réduisent efficacement les erreurs de séquençage, améliorant ainsi la précision des prédictions.

L'évaluation des performances sur des données réelles, validées par des lectures courtes, a montré que RNA-tailor a une bonne sensibilité pour la détection des exons par rapport à FLAIR. Cette validation croisée confirme la fiabilité des prédictions de RNA-tailor. Une analyse exploratoire pour un gène spécifique a démontré l'impact significatif des méthodes de raffinement sur la résolution des isoformes. Les étapes de pré-filtrage, de réalignement et de lissage implémentées dans RNA-tailor contribuent à améliorer la précision des alignements et des prédictions des isoformes.

Bien que RNA-tailor ait montré des performances prometteuses, plusieurs défis et limitations subsistent. L'analyse des données réelles reste complexe en raison de la variabilité des profils d'expression et des artefacts de séquençage. Une validation systématique avec des jeux de données diversifiés est nécessaire pour confirmer la généralisation des résultats. Même si les isoformes fragiles identifiés par RNA-tailor semblent contenir des informations pertinentes, ils nécessitent une validation expérimentale supplémentaire pour confirmer leur existence et leur fonctionnalité.

Chapitre 5

Perspectives et Conclusion

5.1 Perspectives pour le développement de RNA-tailor

La version actuelle de RNA-tailor constitue une première brique stable et performante de la méthode. Cependant, il existe des pistes pour la suite de son développement. En effet, l'amélioration continue des algorithmes d'alignement et de correction est un levier important d'amélioration pour RNA-tailor. Si ces outils sont un jour surpassés en performance par un nouvel état de l'art, ces composantes de RNA-tailor pourront être mises à jour. Différentes perspectives sont étudiées à partir du pipeline de RNA-tailor pour diversifier et enrichir ses capacités d'analyse. Il est envisagé de faire varier le type de séquences de référence, en utilisant par exemple les séquences d'ARNm, de protéines ou les séquences orthologues d'espèces phylogénétiquement proches. L'objectif est d'étudier la possibilité d'identifier de nouveaux isoformes. En effet, l'utilisation d'une séquence de nature différente change la problématique. L'utilisation de la séquence d'un ARNm donne une approche différente qu'une séquence génomique. L'enjeu n'est plus de retrouver les jonctions exoniques correctes mais plutôt d'identifier des variations de longueurs d'exons. Cela permet aussi de s'affranchir des problématiques d'alignement en bordure d'exon sur la référence génomique. Néanmoins, l'identification des événements d'épissage resterait compliqué, comme celle d'un NAGNAG qui poserait le problème de l'identification d'une petite variation de séquence entre deux ARNm. Une autre possibilité serait d'utiliser un ensemble de séquence d'ARNm, de référence ou non. Par rapport à l'utilisation d'un seul ARNm, un ensemble permet une meilleure sélection des lectures. L'utilisation d'un ensemble d'ARNm résout le problème des mécanismes d'épissage difficilement identifiables avec un unique ARNm. Ici la diversité des ARNm orthologues issus d'un même gène mais différemment epissés peut être utile pour la découverte de nouveaux variants, en formant des tuteurs d'alignement précis. Dans le même ordre d'idée, utiliser une séquence protéique apporte une information complémentaire du point de vue biologique : si une séquence ARNm épissée correspond à une protéine, alors cette version de l'ARNm est traduite. De même qu'avec les séquences d'ARNm, les séquences protéiques intègrent l'épissage dans leur séquence. L'enjeu n'est plus de découvrir les jonctions exoniques, mais simplement les mécanismes d'épissage. L'utilisation d'une ou des séquences protéiques nécessite des précautions. Si aucune séquence ARNm ne correspond parfaitement à une protéine, cela peut vouloir dire que ce transcrit n'est pas traduit ou que la protéine subit des modifications post-traductionnelles. Ce type d'analyse peut également être fait après une première passe de prédiction pas RNA-tailor. On peut alors utiliser la séquence d'une protéine prédite par une première analyse. L'intérêt de prendre cette séquence en entrée serait d'observer les différences de sélection des reads à partir de cette séquence protéique simulée. Sa simulation serait faite à partir de l'ORF la plus grande trouvée dans un isoforme ARN prédit.

5.2 Conclusion

L'objectif principal de recherche de cette thèse a été de créer, implémenter puis évaluer les moyens nécessaires pour être capable d'identifier pour un gène, l'ensemble de ces isoformes d'épissage, afin répondre à notre problématique de recherche : *comment être capable d'identifier à partir d'une expérience de transcriptomique de troisième génération et pour un gène donné l'ensemble des ARNm variants d'épissage que ce gène peut produire.* La poursuite de cet objectif s'est cristallisée dans le développement de RNA-tailor, un pipeline d'analyse pour l'identification des isoformes alternatifs d'épissage à partir de séquençage de 3ème génération. Cet outil innove par sa méthodologie et son placement à l'échelle du gène. Cela permet de s'affranchir des contraintes de complexité des méthodes fonctionnant à l'échelle du génome et donc d'utiliser des algorithmes plus consommateur en temps mais plus précis comme l'alignement « est2genome » de exonerate. La philosophie de RNA-tailor est également de laisser la place à la recherche exploratoire de transcript. Pour cela, la méthode est davantage axée vers la sensibilité pour laisser à l'utilisateur le loisir de parcourir les structures d'épissage dans un fichier XLSX facile à prendre en main. RNA-tailor se compose de plusieurs étapes clés : sélection des lectures, filtrage des lectures contaminantes, correction des erreurs, alignement des lectures corrigées, raffinement de la structure des alignements et détermination des isoformes. Le pipeline méthode repose sur l'intégration d'outils publiés tels que isONcorrect pour l'autocorrection des erreurs et exonerate pour l'alignement, minimap2 ou megablast pour la sélection des lectures. Il est également composé de méthodes propres à RNA-tailor et adaptées à l'analyse à l'échelle du gène comme les passes successives de filtrage des lectures, les étapes de raffinement de l'alignement et son module de post-traitement des données. C'est l'utilisation conjointe de l'ensemble de ces méthodes qui rend RNA-tailor innovant et performant. Son pipeline commence par une sélection des lectures à partir de la séquence du gène d'intérêt ou du génome complet de l'espèce, en utilisant des seuils de couverture pour éliminer les faux positifs. Les lectures sélectionnées peuvent être corrigées avec isONcorrect, puis alignées avec exonerate en mode est2genome. Un pré-filtrage élimine les lectures trop différentes en termes de structure, tout en conservant la possibilité d'un rachat par le réalignement afin de corriger les potentielles erreurs d'alignement. Enfin, on applique la méthode du lissage des bordures pour consolider les jonctions exoniques. Les lectures corrigées et alignées sont regroupées en isoformes selon leurs structures introniques. Enfin, il est possible d'utiliser les méthodes de post-traitement des données pour explorer la nature des isoformes prédits et affiner les prédictions.

Les résultats du benchmark réalisés montrent que RNA-tailor possède une bonne sensibilité et précision dans la détection des isoformes, surpassant souvent FLAIR dans la reconnaissance des événements d'épissage alternatif tels que la détection de rétention d'intron et d'exon cassette. Il a également démontré une robustesse face aux erreurs de séquençage, maintenant une sensibilité élevée même à des taux d'erreur élevés. Les résultats obtenus avec RNA-tailor ont été validés en les comparant à ceux de FLAIR et Freddie, en utilisant des jeux de données réelles et simulées. La validation croisée avec des lectures courtes a permis de confirmer la fiabilité des prédictions de chacun des outils.

Pour améliorer RNA-tailor, il est envisagé de continuer à perfectionner les algorithmes de correction et de veiller à l'évolution de l'état de l'art. Il est envisageable de continuer le développement de RNA-tailor pour passer son échelle d'analyse à celle du génome. La singularité de RNA-tailor peut également être exploitée pour mener d'autres types d'analyses innovantes. Par exemple, l'utilisation de séquences de référence variées, telles que les séquences d'ARNm et de protéines, pour identifier de nouveaux isoformes est envisageable. L'intégration de séquences orthologues d'espèces proches pourrait avoir un intérêt pour l'analyse d'espèces non-modèles.

Les principaux résultats proposés dans ce travail de thèse sont les suivants :

— RNA-tailor est notre proposition de réponse à la problématique de recherche sur la résolution de l'inventaire des transcrits variants d'épissage à l'échelle d'un gène et à partir de données de séquençage 3ème génération.
— Son approche est singulière comparée aux autres outils d'identification des isoformes de part son échelle d'analyse au niveau du gène qui lui permet l'utilisation de méthode précise comme exonerate en mode « est2genome » et local, d'exporter une représentation claire de la structure d'alignement au format XLSX adapté à l'étude exploratoire des mécanisme d'épissage.
— RNA-tailor propose des performances à l'état de l'art en termes de sensibilité et de précision qui surpassent les outils qui lui sont comparés dans le benchmark.
— Son module de post-traitement permet d'étudier la nature et la composition des transcrits pour réaliser du « *fine tuning* » propre aux besoins de l'utilisateur.
— Cependant, la résolution de l'inventaire des isoformes alternatifs est une tâche complexe et notre travail a mis en évidence que les outils de prédiction des isoformes ne sont pas encore arrivés à un consensus sur les données réelles.

Bibliographie

[NW70] S. B. Needleman et C. D. Wunsch. « A general method applicable to the search for similarities in the amino acid sequence of two proteins ». eng. In : *Journal of Molecular Biology* 48.3 (mars 1970), p. 443-453. issn : 0022-2836. doi : 10.1016/0022-2836(70)90057-4.

[Alt+90] S. F. Altschul et al. « Basic local alignment search tool ». eng. In : *Journal of Molecular Biology* 215.3 (oct. 1990), p. 403-410. issn : 0022-2836. doi : 10.1016/S0022-2836(05)80360-2.

[Lan+01] Eric S. Lander et al. « Initial sequencing and analysis of the human genome ». en. In : *Nature* 409.6822 (fév. 2001), p. 860-921. issn : 1476-4687. doi : 10.1038/35057062.

[ML02] Barmak Modrek et Christopher Lee. « A genomic view of alternative splicing ». In : *Nature Genetics* 30.1 (jan. 2002), p. 13-19. issn : 1061-4036. doi : 10.1038/ng0102-13.

[SB05] Guy St C. Slater et Ewan Birney. « Automated generation of heuristics for biological sequence comparison ». en. In : *BMC Bioinformatics* 6.1 (fév. 2005), p. 31. issn : 1471-2105. doi : 10.1186/1471-2105-6-31. url : https://doi.org/10.1186/1471-2105-6-31 (visité le 17/04/2023).

[Pen+08] Tao Peng et al. « Functional importance of different patterns of correlation between adjacent cassette exons in human and mouse ». In : *BMC Genomics* 9.1 (avr. 2008), p. 191. issn : 1471-2164. doi : 10.1186/1471-2164-9-191.

[Wan+08] Eric T. Wang et al. « Alternative isoform regulation in human tissue transcriptomes ». en. In : *Nature* 456.7221 (nov. 2008), p. 470-476. issn : 1476-4687. doi : 10.1038/nature07509.

[Cam+09] Christiam Camacho et al. « BLAST+ : architecture and applications ». In : *BMC Bioinformatics* 10.1 (déc. 2009), p. 421. issn : 1471-2105. doi : 10.1186/1471-2105-10-421.

[WWL09] Markus C. Wahl, Cindy L. Will et Reinhard Lührmann. « The Spliceosome : Design Principles of a Dynamic RNP Machine ». In : *Cell* 136.4 (fév. 2009), p. 701-718. issn : 0092-8674. doi : 10.1016/j.cell.2009.02.009.

[FWH10] Martin C. Frith, Raymond Wan et Paul Horton. « Incorporating sequence quality data into alignment improves DNA read mapping ». In : *Nucleic Acids Research* 38.7 (avr. 2010), e100. issn : 0305-1048. doi : 10.1093/nar/gkq010.

[KLA10] Hadas Keren, Galit Lev-Maor et Gil Ast. « Alternative splicing and evolution : diversification, exon definition and function ». en. In : *Nature Reviews Genetics* 11.5 (mai 2010), p. 345-355. issn : 1471-0064. doi : 10.1038/nrg2776.

[Bra+12] Robert K. Bradley et al. « Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution ». en. In : *PLOS Biology* 10.1 (jan. 2012), e1001229. issn : 1545-7885. doi : 10.1371/journal.pbio.1001229.

[Heg+12] Anna Hegele et al. « Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome ». In : *Molecular Cell* 45.4 (fév. 2012), p. 567-580. ISSN : 1097-2765. DOI : 10.1016/j.molcel.2011.12.034.

[Zim+13] Aleksey V. Zimin et al. « The MaSuRCA genome assembler ». In : *Bioinformatics* 29.21 (nov. 2013), p. 2669-2677. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btt476.

[RA15] Anthony Rhoads et Kin Fai Au. « PacBio Sequencing and Its Applications ». In : *Genomics, Proteomics & Bioinformatics* 13.5 (oct. 2015), p. 278-289. ISSN : 1672-0229. DOI : 10.1016/j.gpb.2015.08.002.

[ŁTD16] Anna Łabno, Rafał Tomecki et Andrzej Dziembowski. « Cytoplasmic RNA decay pathways - Enzymes and mechanisms ». In : *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1863.12 (déc. 2016), p. 3125-3147. ISSN : 0167-4889. DOI : 10.1016/j.bbamcr.2016.09.023.

[BG17] Francisco E. Baralle et Jimena Giudice. « Alternative splicing as a regulator of development and tissue identity ». In : *Nature Reviews Molecular Cell Biology* 18.7 (juill. 2017), p. 437-451. ISSN : 1471-0072. DOI : 10.1038/nrm.2017.27.

[MRH17] Francesco P. Marchese, Ivan Raimondi et Maite Huarte. « The multidimensional mechanisms of long noncoding RNA function ». In : *Genome Biology* 18.1 (oct. 2017), p. 206. ISSN : 1474-760X. DOI : 10.1186/s13059-017-1348-2.

[Vas+17] Robert Vaser et al. « Fast and accurate de novo genome assembly from long uncorrected reads ». In : *Genome research* 27.5 (2017), p. 737-746.

[Yan+17] Chen Yang et al. « NanoSim : nanopore sequence read simulator based on statistical characterization ». In : *GigaScience* 6.4 (avr. 2017), gix010. ISSN : 2047-217X. DOI : 10.1093/gigascience/gix010.

[Li18] Heng Li. « Minimap2 : pairwise alignment for nucleotide sequences ». In : *Bioinformatics* 34.18 (sept. 2018), p. 3094-3100. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bty191. URL : https://doi.org/10.1093/bioinformatics/bty191 (visité le 17/04/2023).

[SK18] Hajime Suzuki et Masahiro Kasahara. « Introducing difference recurrence relations for faster semi-global alignment of long sequences ». In : *BMC Bioinformatics* 19.1 (fév. 2018), p. 45. ISSN : 1471-2105. DOI : 10.1186/s12859-018-2014-8.

[Bus+19] Elena Bushmanova et al. « rnaSPAdes : a de novo transcriptome assembler and its application to RNA-Seq data ». In : *GigaScience* 8.9 (sept. 2019), giz100. ISSN : 2047-217X. DOI : 10.1093/gigascience/giz100.

[Byr+19] Ashley Byrne et al. « Realizing the potential of full-length transcriptome sequencing ». In : *Philosophical Transactions of the Royal Society B* 374.1786 (2019), p. 20190097. ISSN : 0962-8436. DOI : 10.1098/rstb.2019.0097.

[Kov+19] Sam Kovaka et al. « Transcriptome assembly from long-read RNA-seq alignments with StringTie2 ». en. In : *Genome Biology* 20.1 (déc. 2019), p. 278. ISSN : 1474-760X. DOI : 10.1186/s13059-019-1910-1.

[Mon+19] Geoffray Monteuuis et al. « The changing paradigm of intron retention : regulation, ramifications and recipes ». In : *Nucleic Acids Research* 47.22 (déc. 2019), p. 11497-11513. ISSN : 0305-1048. DOI : 10.1093/nar/gkz1068.

[RQQ19] Maxime Rotival, Hélène Quach et Lluis Quintana-Murci. « Defining the genetic and evolutionary architecture of alternative splicing in response to infection ». In : *Nature Communications* 10.1 (avr. 2019), p. 1671. DOI : 10.1038/s41467-019-09689-7.

[Ses+19] Camille Sessegolo et al. « Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules ». In : *Scientific Reports* 9.1 (oct. 2019), p. 14908. doi : 10.1038/s41598-019-51470-9.

[Wic19] Ryan R. Wick. « Badread : simulation of error-prone long reads ». en. In : *Journal of Open Source Software* 4.36 (avr. 2019), p. 1316. issn : 2475-9066. doi : 10.21105/joss.01316.

[Bes+20] Cláudia Bessa et al. « Alternative Splicing : Expanding the Landscape of Cancer Biomarkers and Therapeutics ». en. In : *International Journal of Molecular Sciences* 21.2323 (jan. 2020), p. 9032. issn : 1422-0067. doi : 10.3390/ijms21239032.

[Gre+20] Immanuel D Green et al. « Macrophage development and activation involve coordinated intron retention in key inflammatory regulators ». In : *Nucleic Acids Research* 48.12 (juill. 2020), p. 6513-6529. issn : 0305-1048. doi : 10.1093/nar/gkaa435.

[Haf+20] Saber Hafezqorani et al. « Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data ». In : *GigaScience* 9.6 (juin 2020), giaa061. issn : 2047-217X. doi : 10.1093/gigascience/giaa061.

[LVE20] Glennis A. Logsdon, Mitchell R. Vollger et Evan E. Eichler. « Long-read human genome sequencing and its applications ». en. In : *Nature Reviews Genetics* 21.10 (oct. 2020), p. 597-614. issn : 1471-0064. doi : 10.1038/s41576-020-0236-x.

[Nip+20] Ka Ming Nip et al. « RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes ». In : *Genome Research* 30.8 (août 2020), p. 1191-1200. issn : 1088-9051. doi : 10.1101/gr.260174.119.

[Oku+20] Mariko Okubo et al. « Exon skipping induced by nonsense/frameshift mutations in DMD gene results in Becker muscular dystrophy ». en. In : *Human Genetics* 139.2 (fév. 2020), p. 247-255. issn : 1432-1203. doi : 10.1007/s00439-019-02107-4.

[PP20] Geo Pertea et Mihaela Pertea. « GFF Utilities : GffRead and GffCompare ». In : *F1000Research* 9 (2020), ISCB Comm J-304. doi : 10.12688/f1000research.23297.2.

[Pre+20] Marco Preussner et al. « Splicing-accessible coding $3_UTRscontrolproteinstabilityandinteracti$ In : *Genome Biology* 21.1 (juill. 2020), p. 186. issn : 1474-760X. doi : 10.1186/s13059-020-02102-3.

[SM20] Kristoffer Sahlin et Paul Medvedev. « De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality Value-Based Algorithm ». In : *Journal of Computational Biology* 27.4 (avr. 2020), p. 472-484. doi : 10.1089/cmb.2019.0299.

[Sch+20] Katharina Schwarze et al. « The complete costs of genome sequencing : a microcosting study in cancer and rare diseases from a single center in the United Kingdom ». In : *Genetics in Medicine* 22.1 (jan. 2020), p. 85-94. issn : 1098-3600. doi : 10.1038/s41436-019-0618-7.

[Tan+20] Alison D. Tang et al. « Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns ». en. In : *Nature Communications* 11.1 (mars 2020), p. 1438. issn : 2041-1723. doi : 10.1038/s41467-020-15171-6. url : https://www.nature.com/articles/s41467-020-15171-6 (visité le 17/04/2023).

[Cho21] Hilbert Lam Chokyotager. « Pypi orffinder ». In : *https ://pypi.org/project/orffinder/* (2021).

[Kan+21] Nisha KANWAR et al. « PacBio sequencing output increased through uniform and directional fivefold concatenation ». In : *Scientific Reports* 11.1 (sept. 2021), p. 18065. DOI : 10.1038/s41598-021-96829-z.

[Lam+21] Su Datt LAM et al. « Biological impact of mutually exclusive exon switching ». en. In : *PLOS Computational Biology* 17.3 (mars 2021), e1008708. ISSN : 1553-7358. DOI : 10.1371/journal.pcbi.1008708.

[Sah+21] Kristoffer SAHLIN et al. « Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis ». In : *Nature Communications* 12.1 (2021), p. 2. DOI : 10.1038/s41467-020-20340-8.

[SN21] Nicholas STOLER et Anton NEKRUTENKO. « Sequencing error profiles of Illumina sequencing instruments ». In : *NAR Genomics and Bioinformatics* 3.1 (mars 2021), lqab019. ISSN : 2631-9268. DOI : 10.1093/nargab/lqab019.

[Tan+21] Zhichao TANG et al. « RNA-Targeting Splicing Modifiers : Drug Development and Screening Assays ». en. In : *Molecules* 26.88 (jan. 2021), p. 2263. ISSN : 1420-3049. DOI : 10.3390/molecules26082263.

[Wan+21] Yunhao WANG et al. « Nanopore sequencing technology, bioinformatics and applications ». en. In : *Nature Biotechnology* 39.11 (nov. 2021), p. 1348-1365. ISSN : 1546-1696. DOI : 10.1038/s41587-021-01108-x. URL : https://www.nature.com/articles/s41587-021-01108-x (visité le 17/04/2023).

[Mil+22] Rachel MILLER et al. « Enhanced protein isoform characterization through long-read proteogenomics ». In : *Genome Biology* 23 (mars 2022). DOI : 10.1186/s13059-022-02624-y.

[OHA22] Yukiteru ONO, Michiaki HAMADA et Kiyoshi ASAI. « PBSIM3 : a simulator for all types of PacBio and ONT long reads ». In : *NAR Genomics and Bioinformatics* 4.4 (déc. 2022), lqac092. ISSN : 2631-9268. DOI : 10.1093/nargab/lqac092. URL : https://doi.org/10.1093/nargab/lqac092 (visité le 05/06/2024).

[Che+23] Ying CHEN et al. « Context-aware transcript quantification from long-read RNA-seq data with Bambu ». In : *Nature Methods* 20.8 (2023), p. 1187-1195. ISSN : 1548-7091. DOI : 10.1038/s41592-023-01908-w.

[Cho+23] Karine CHOQUET et al. « Pre-mRNA splicing order is predetermined and maintains splicing fidelity across multi-intronic transcripts ». In : *Nature Structural & Molecular Biology* 30.8 (août 2023), p. 1064-1076. ISSN : 1545-9993. DOI : 10.1038/s41594-023-01035-2.

[Gao+23] Yuan GAO et al. « ESPRESSO : Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data ». In : *Science Advances* 9.3 (2023), eabq5072. DOI : 10.1126/sciadv.abq5072.

[Ora+23] Baraa ORABI et al. « Freddie : annotation-independent detection and discovery of transcriptomic alternative splicing isoforms using long-read sequencing ». In : *Nucleic Acids Research* 51.2 (jan. 2023), e11. ISSN : 0305-1048. DOI : 10.1093/nar/gkac1112. URL : https://doi.org/10.1093/nar/gkac1112 (visité le 17/04/2023).

[PS23] Alexander J PETRI et Kristoffer SAHLIN. « isONform : reference-free transcriptome reconstruction from Oxford Nanopore data ». In : *Bioinformatics* 39.Supplement_1 (juin 2023), p. i222-i231. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btad264.

[Prj+23] Andrey D. PRJIBELSKI et al. « Accurate isoform discovery with IsoQuant using long reads ». en. In : *Nature Biotechnology* (jan. 2023), p. 1-4. ISSN : 1546-1696. DOI : 10.1038/s41587-022-01565-y.

[Sad+23] Harisankar SADASIVAN et al. « Accelerating Minimap2 for Accurate Long Read Alignment on GPUs ». In : *Journal of Biotechnology and Biomedicine* 06 (jan. 2023). DOI : 10.26502/jbb.2642-91280067.

[SD23] Ewa SYBILSKA et Agata DASZKOWSKA-GOLEC. « Alternative splicing in ABA signaling during seed germination ». In : *Frontiers in Plant Science* 14 (mars 2023), p. 1144990. ISSN : 1664-462X. DOI : 10.3389/fpls.2023.1144990.

[Yan+23] Chen YANG et al. « Characterization and simulation of metagenomic nanopore sequencing data with Meta-NanoSim ». In : *GigaScience* 12 (jan. 2023), giad013. ISSN : 2047-217X. DOI : 10.1093/gigascience/giad013.

[Liu+24] Wang LIU-WEI et al. « Sequencing accuracy and systematic errors of nanopore direct RNA sequencing ». In : *BMC Genomics* 25.1 (mai 2024), p. 528. ISSN : 1471-2164. DOI : 10.1186/s12864-024-10440-w.

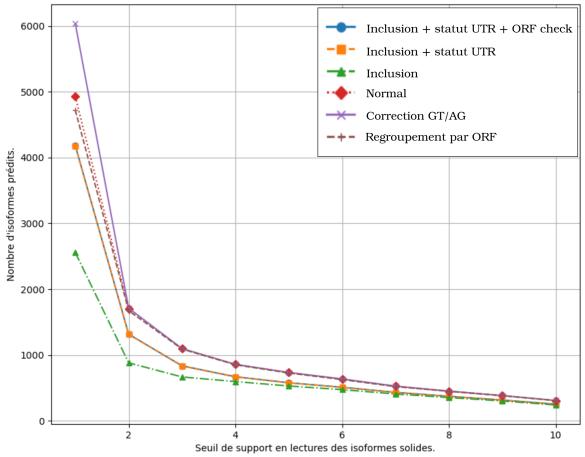
[Par+24] Francisco J. PARDO-PALACIOS et al. « SQANTI3 : curation of long-read transcriptomes for accurate identification of known and novel isoforms ». en. In : *Nature Methods* 21.5 (mai 2024), p. 793-797. ISSN : 1548-7105. DOI : 10.1038/s41592-024-02229-2.

Annexe A

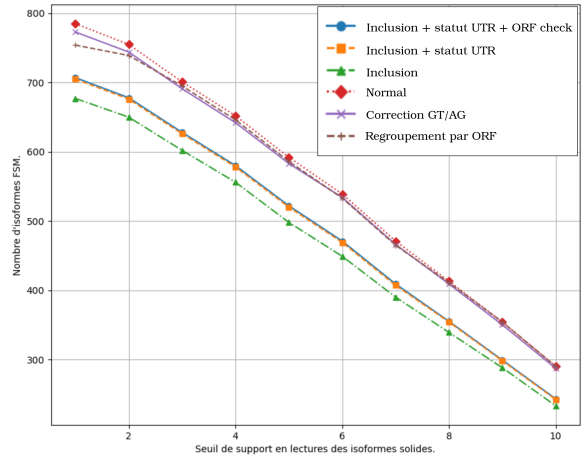
Annexe

A.1 Effet du filtre de support de lecture

Nous avons voulu observer l'effet de la catégorisation des isoformes prédits en solide et fragile en fonction du support en nombre de lectures. D'après le graphique A.1 a), le



(a)



(b)

FIGURE A.1 – Évolution du nombre total d'isoformes (b) et du nombre de FSM (b) en fonction du seuil de support pour les différentes stratégies.

nombre d'isoforme varie nettement en fonction du seuil de support de lecture demandé pour définir les isoformes solides. Le nombre de isoforme prédit chute fortement dès la valeur deux de support alors que le nombre FSM diminue de façon linéaire, ce qui est attendu, car le seuil de support est appliqué après les méthodes de regroupement. A partir d'une valeur de support de quatre, l'effet des méthodes de regroupement/filtrage des isoformes est effacé par le seuil de support. Toujours d'après ces deux graphiques, le choix d'un seuil de support des lectures à deux est un filtre efficace pour réduire le nombre d'isoformes et maximiser l'effet des méthodes de regroupement. La mise en perspective de ces résultats avec celui de la variation du nombre de isoforme FSM en fonction du seuil de support en lecture, en b), permet de constater que deux est effectivement un seuil intéressant pour éliminer un grand nombre d'isoforme fragile, tout en gardant une très haute sensibilité au FSM.

A.2 QR code vers le dépôt git de RNA-tailor

FIGURE A.2 – QR code vers le dépôt git de RNA-tailor.