Développement de méthodes bio-informatiques pour l'étude de l'épissage chez les espèces non modèles : épissage complexe et apport des technologies de séquençage de 3eme génération

Camille Sessegolo

N° d'ordre NNT : 2021LYSE1218

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de :

l'Université Claude Bernard Lyon 1

Ecole Doctorale 341
Évolution Écosystèmes Microbiologie Modélisation

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 22/10/2021, par :

Camille Sessegolo

Développement de méthodes bio-informatiques pour l'étude de l'épissage chez les espèces non modèles : Épissage complexe et apport des technologies de séquençage de 3eme génération.

Devant le jury composé de :

VIEIRA-HEDDI Cristina, Professeure des universités, UCBL1	Présidente
FERNANDEZ DE LUCO Reini, Chargée de Recherche, IGH	Rapporteure
TOUZET Hélène, Directrice de Recherche, CNRS	Rapporteure
NAFFAKH Nadia, Directrice de Recherche, Institut Pasteur, Paris	Examinateur.rice
LACROIX Vincent, Maître de Conférences, UCBL1	Directeur de thèse
MARY Arnaud , Maître de Conférences, UCBL1	Co-Directeur de thèse
PETERLONGO Pierre, Chargé de Recherche, INRIA Rennes	Invité
AURY Jean-Marc, Ingénieur de Recherche, Génoscope, CEA d'Évry	Invité

Université Claude Bernard – LYON 1

Président de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALLIER
Vice-Président de la Commission de Recherche	M. Petru MIRONESCU
Directeur Général des Services	M. Pierre ROLLAND

COMPOSANTES SANTE

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directrice : Mme Christine VINCIGUERRA

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Administrateur Provisoire : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

Résumé

Les gènes des organismes eucaryotes sont structurés en exons et en introns. Lors de l'épissage, les introns sont retirés et les exons reliés entre eux. L'utilisation des sites d'épissage par la machinerie cellulaire peut varier d'un transcrit à l'autre pour un même gène. L'épissage alternatif permet alors à un seul gène de produire plusieurs transcrits et parfois plusieurs protéines. L'étude des données issues du séquençage des transcrits (RNAseq) nous permet d'étudier l'épissage. Actuellement deux technologies de séquençage coexistent : les technologies de seconde génération, permettant de produire des lectures courtes (100 à 250pb) avec un taux d'erreur faible et les technologies de 3ème génération permettant de produire des lectures longues (plusieurs kb) avec des taux d'erreur plus élevés. Dans un premier temps, j'ai analysé des jeux de données Nanopore (lectures longues) afin de comprendre comment ces technologies, récentes et en constante évolution, peuvent nous aider à étudier les transcriptomes eucaryotes. Plus particulièrement, je me suis demandée si les quantifications des gènes et des transcrits obtenues étaient fiables. L'utilisation de spike-in -transcrits artificiels dont on connaît la quantification- nous a permis de montrer que, parmi les différents protocoles testés, les quantifications obtenues avec le protocole RNA direct sont les plus fiables. De plus, contrairement à ce à quoi l'on s'attendait, les lectures ne couvrent pas systématiquement des transcrits complets. Ensuite, je me suis intéressée à la modélisation des évènements d'épissage alternatif complexes chez les espèces non modèles. L'assembleur local de transcriptome, KisSplice [64], développé dans l'équipe compare deux à deux les transcrits, même lorsqu'il y a plus de deux transcrits localement. Je propose ici une nouvelle échelle d'étude de l'épissage qui permet de considérer toutes les variations d'épissages observées entre deux exons constitutifs à tous les transcrits d'un gène.

Mots-Clefs : *Épissage alternatif, RNAseq, espèce non modèle, lectures longues, transcriptomique, bioinformatique.*

Abstract

Eukaryotic genes are composed of exons and introns. Introns are spliced out during the maturation of pre-mRNA to mRNA. Splice site usage may vary from one transcript to another for a same gene. Alternative splicing is a source of diversity in eucaryotic transcriptomes and one gene can sometimes lead to several proteins. We can study alternative splicing using RNAseq data. Nowadays second (2GS) and third generation sequencing (3GS) technologies coexist. 2GS produce short (from 100 to 250 pb) high quality reads whereas 3GS produce long reads (up to several kilobases) but with a lot of errors. In the first part of my PhD, I analysed Nanopore long reads datasets to understand how this technology can help us to study eucaryotic transcriptomes. Particularly, I wondered if transcripts and genes quantifications obtained with Nanopore data were reliable. We used Spike-in (artificial transcripts from which we know the real quantification) and we showed that the most precise quantifications were obtained with the RNA direct protocol. Furthermore, we observed that only a fraction of the long reads covered full length transcripts. Then, I worked on a new model for complex alternative splicing events in non-model species. KisSplice [64], the local RNAseq assembler developped in the team, always considers parwise event even when there are more than two transcripts locally. I propose here a new scale to study alternative splicing : we consider all the splicing variations observed between two constitutive exons of a gene.

key-Words : *Alternative splicing, RNAseq, non-model species, long reads, transcriptomic, bioinformatic*

Remerciements

Je tiens tout d'abord à remercier les membres du jury : Cristina VIEIRA-HEDDI, Reini FERNANDEZ DE LUCO, Hélène TOUZET, Nadia NAFFAKH, Pierre PETERLONGO et Jean-Marc AURY d'avoir accepté d'assister à ma soutenance et tout particulièrement Reini FERNANDEZ DE LUCO et Hélène TOUZET pour le temps qu'elles ont accordé à la lecture et à l'évaluation de mon manuscrit.

Je souhaitais aussi remercier les membres de mon comité de suivi, qui m'ont éclairé plus d'une fois sur des points importants et toujours avec bienveillance : merci Laurent DURET, Laurent JACOB, Mathieu GABUT, et Tristan LEFEBURE.

Je voudrais aussi remercier tous les membres de l'équipe BAOBAB (actuels ou passés) avec qui j'ai eu le plaisir d'interagir : Alex, Antoine, Audric, Blerina, Camille, Carol, Claire, Hélène, Hélio, Hermes, Irene, Laura, Laurent, Louis, Mariana, Marianne, Martin, Mattia, Nina et Taneli. Merci Marie-France de m'avoir permis de faire ma thèse dans un environnement aussi riche humainement et aussi bienveillant. Merci Clara pour toutes les pauses thé passées ensemble et surtout pour ton soutien infaillible. Merci Leandro pour tous les thés / cafés / verres de vins au lait concentré sucré pris ensemble, pour les fous rires dans le bureau et pour avoir gardé ton âme d'enfant qui m'a plus d'une fois aidé à garder le moral. Merci Ricardo pour tous les bons moments passés ensemble, pour tes conseils et pour la recette de la pâte à pizza. Merci Éric et Nicolas pour toutes les discussions et les pauses déjeuners salvatrices au labo ces derniers temps.

Je souhaite remercier tout particulièrement mes directeurs de thèse, Vincent LACROIX, Arnaud MARY et Jean Marc AURY pour leur soutien et leur bienveillance. Merci Vincent de m'avoir supporté - dans tous les sens du terme - toutes ces années. Tu m'as poussé à avancer même lorsque je n'y croyais plus, tu m'as aussi dit que j'allais survivre dans un moment où

j'avais besoin de l'entendre et tu avais raison. Tu as cru en moi et j'ai beaucoup grandi grâce à toi. Je ne sais pas comment te remercier pour tout ça, ce sera au moins écrit ici. Merci pour tout.

Merci à mes amis pour leur présence, et les moments partagés. Merci Vincent M., Benoit, et Insun. Merci Alex pour toutes ces années passées ensemble, les récréations à discuter plutôt qu'à jouer au foot, les exos de maths, les dissert de philo et les versions de grec anciens, les premières bières en Irlande, les soirées à "la traboule", les après-midi jeux, et les moments avec ton filleul. Grandir avec toi m'a permis de devenir qui je suis maintenant mais aussi de développer ma curiosité scientifique depuis toute petite et je n'en serai sûrement pas là sans toi. Merci Julie de me démontrer que l'on peu vivre de ses rêves.

Merci Anne-Marie pour les prises de conscience successives.

Enfin, je souhaite remercier ma famille, David, Côme et Nayane. Merci Nayane pour ta douceur et tes clins d'yeux. Merci David pour ton amour et ton soutien dans toutes les épreuves traversées ensemble ces douze dernières années (y compris ma thèse). Merci Côme d'avoir fait de moi une maman et pour tout ce que tu m'a appris ces 20 derniers mois. Merci mon tout petit bout pour ta tendresse, ton enthousiasme et tes yeux qui brillent. Sans vous trois, rien de tout cela n'aurait de sens.

***"Les amis, c'est une famille dont on a choisi les membres."* Jean Baptiste Alphonse Karr, Les guêpes (1847)**

Table des matières

I Introduction **19**

1 Support de l'information génétique **21**

1.1 Structure moléculaire 21

1.2 Dogme central de la biologie moléculaire 22

2 Épissage **25**

2.1 Définition générale 25

2.2 Mécanismes moléculaires 26

2.2.1 Réaction d'épissage 26

2.2.2 Reconnaissance des exons 27

2.3 Épissage alternatif 28

2.3.1 Définition 28

2.3.2 Les différents types d'évènements d'épissage alternatif 29

2.3.3 Prévalence de l'épissage alternatif 30

2.4 Régulation de l'épissage alternatif 31

2.4.1 Erreur de la machinerie d'épissage et régulation 31

2.4.2 Épissage et NMD 32

3 Séquençage du transcriptome **35**

3.1 Historique 35

3.2 Séquençage de 2eme génération 36
3.2.1 Préparation des librairies 36
3.2.2 Séquençage 37
3.2.3 Caractéristiques des données 38
3.3 Séquençage de 3eme génération 39
3.3.1 PacBio 40
3.3.2 Oxford Nanopore 41

4 Analyse de l'épissage à partir des données RNAseq 47
4.1 Analyse des données de 2ème génération 47
4.1.1 Méthodes basées sur l'alignement 48
4.1.2 Méthodes basées sur l'assemblage 51
4.1.3 Échelle d'étude 51
4.1.4 Identification des événements d'épissage/des transcrits 52
4.1.5 Analyse différentielle 54
4.2 Analyse des données de 3eme génération. 57

II Objectifs de la thèse 59

III Apport des données de troisième génération pour l'étude des transcriptomes 63

1 Contexte 65

2 Publication dans Scientific Reports (2019) 69

3 Matériel supplémentaire à la publication précédente 83

4 Perspectives 97
4.1 Résolution des problèmes de mapping multiple sur les transcrits 97

4.2 Évolution de la profondeur et des taux d'erreurs 99

IV Modélisation des évènements d'épissage complexes 101

1 Background 103

1.1 Méthodes bio-informatiques existantes 103

1.2 KisSplice et Graphes de De Bruijn 104

1.2.1 Construction du graphe de De Bruijn (DBG) 104

1.2.2 Détection des événements d'épissage 104

1.3 Évènements d'épissage complexe 105

1.4 Régulation des évènements d'épissage complexe 107

2 Pipeline d'analyse des événements complexes 109

2.1 Présentation générale 109

2.2 Définition des bulles complexes 109

2.3 Construction des bulles complexes 112

2.4 Modélisation du choix des sites d'épissage 113

2.5 Quantification des bulles complexes 114

2.6 Analyse différentielle 115

3 Pertinence de la modélisation 119

3.1 Exemple d'évènements régulés par PTBP1 119

3.1.1 Pbx1 120

3.1.2 Gabbr1 123

3.2 Exemples d'évènements régulés par RED 124

3.2.1 CCNT2 126

3.2.2 CLK1 132

4 Discussion et perspectives 137

4.1 Échelle d'étude intermédiaire pour étudier l'épissage 137

4.2 Puissance statistique 138
4.3 Comparaison de transcrits mineurs entre eux 139
4.4 Questions mécanistiques 140
4.5 Régulation de l'épissage et bruit de la machinerie d'épissage 140
4.6 Perspectives concernant les lectures longues 141

Table des figures

1 Structure des molécules d'ADN et d'ARN 22

2 Dogme central de la biologie moléculaire 23

3 Exemple d'un gène transcrit en pré-ARNm puis épissé. 25

4 Réaction d'épissage simplifiée 27

5 Reconnaissance des exons 28

6 Exemple d'un gène transcrit en pré-ARNm puis épissé alternativement . . . 29

7 Les différents types d'évènements d'épissage simples 30

8 Préparation des librairies Illumina 38

9 Séquençage illumina 39

10 Séquençage par Pacbio 40

11 Préparation des librairies nanopore cDNA 42

12 Préparation de librairies nanopore RNAdirect 43

13 Séquençage nanopore 44

14 Alignement des lectures RNAseq 48

15 Aligneurs *exon first* et *seed and extend* 49

16 Calcul des valeurs de PSI et ΔPSI 56

17 Couverture du SIRV 505 98

18 Cas où l'on peut réattribuer une lecture à un nouveau transcrit 99

19 Graphe de De Bruijn 104
20 Les événements d'épissage forment des structures dans le DBG 106
21 Pipeline d'identification et de quantification des événements d'épissage complexe. 110
22 Séparation des bulles complexes ayant un noeud en commun 111
23 Modélisation du choix des sites d'épissage 114
24 Exemples de calculs des PSI et des ΔPSI 116
25 Exemple du saut de l'exon 7 de Pbx1. 121
26 Sorties de MAJIQ pour Pbx1 123
27 Exemple d'une bulle complexe dans le gène Gabbr1. 125
28 Visualisation du graphe de De Bruijn formé par le gène CCNT2. 127
29 Exemple d'évènement complexe pour le gène CCNT2. 129
30 Exemple du gène CLK1 132

Liste des tableaux

1 Taux d'erreur obtenus pour les différents jeux de données. 100

Première partie

Introduction

1 | Support de l'information génétique

1.1 Structure moléculaire

Chez tous les organismes, qu'ils soient eucaryotes ou procaryotes, l'information génétique est stockée via la macro-molécule d'ADN (Acide désoxyribonucléique). Les travaux de Watson et Crick [82] ainsi que ceux de Rosalind Franklin ont permis de montrer que celle-ci est composée de deux brins complémentaires, structurés en forme de double hélice. D'un point de vue biochimique, chacun des brins est composé d'une chaine de sucre (désoxyribose) reliés entre eux par des liaisons phosphodiester. Chacun de ces sucres porte une base nucléotidique. Quatre bases composent l'ADN : L'Adénine (A), la Guanine (G), la Cytosine (C) et la Thymine (T). Les paires de bases A et T ainsi que G et C sont complémentaires. Ces nucléotides sont donc toujours associés l'un à l'autre au sein de la double hélice d'ADN (Figure 1). Chez les organismes eucaryotes, l'information génétique est stockée dans le noyau. Les doubles brins d'ADN y sont alors, la plupart du temps, fortement condensés.

On s'intéresse dans cette thèse aux gènes, c'est-à-dire aux régions génomiques permettant de coder pour des protéines, mais on sait maintenant qu'une partie importante -en proportion variable selon les organismes- des génomes est composée d'ADN dit non codant. Le dogme central de la biologie moléculaire permet d'établir le lien entre gènes et protéines.

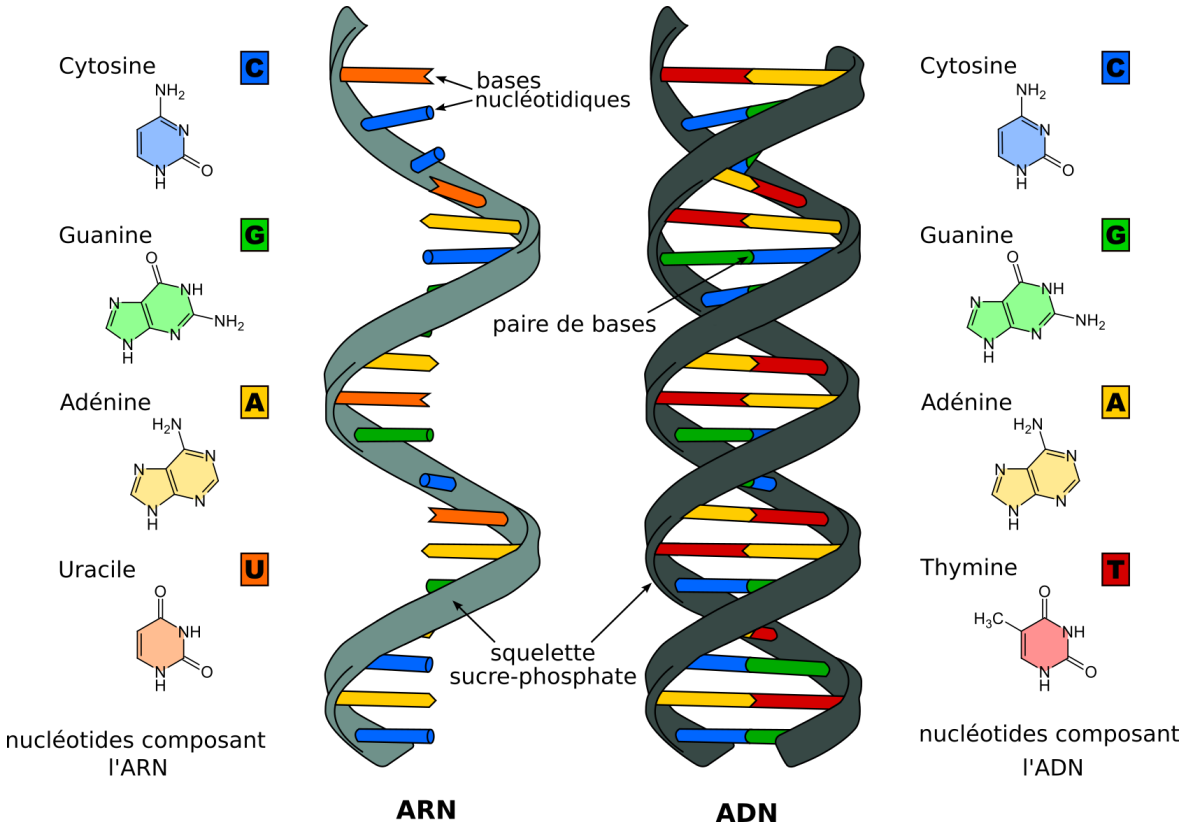


FIGURE 1 – Structure des molécules d'ADN et d'ARN. Figure adaptée depuis https ://commons.wikimedia.org/wiki/RNA.

1.2 Dogme central de la biologie moléculaire

Les gènes permettent de stocker l'information nécessaire à la production des protéines. Pour cela, chacun des gènes exprimé dans une cellule est, tout d'abord, transcrit en ARN messager avant de pouvoir être traduit en protéine (Figure 2).

Lors de la transcription, l'ADN est transcrit en une autre macro-molécule composée d'acide nucléique : l'ARN (Acide désoxyribo-nucléique). Les molécules d'ARN sont simple brin et donc moins stables que l'ADN. Les molécules d'ARN servent en effet d'étape intermédiaire entre les gènes et les protéines, contrairement aux molécules d'ADN, très stables, qui permettent un stockage pérenne de l'information génétique (Figure 1). Les brins d'ARN, sont

composés des bases nucléotidiques suivantes : L'Adénine (A), la Guanine (G), la Cytosine (C) et l'Uracile (U). L'Uracile remplace dans l'ARN les thymines présentes dans l'ADN. La transcription chez les organismes eucaryotes et sa régulation sont des processus complexes impliquant notamment différentes ARN polymérases. L'ARN polymérase ainsi que différents facteurs de transcription se fixent en amont du promoteur du gène et permetttent ainsi de déclencher la transcription du gène à partir du site d'initiation de la transcription. Les molécules d'ARN messager ainsi produites sont ensuite maturées. Plusieurs étapes importantes ont lieu lors de la maturation de l'ARN messager. Une coiffe est ajoutée à l'extrémité 5' des transcrits, et l'extrémité 3' est polyadénylée. L'épissage permet enfin d'obtenir des transcrits matures.

Dans la plupart des cas, les molécules d'ARN matures sont ensuite traduites en protéines (macro molécules composées d'acide aminés). Les bases nucléotidiques, associées trois par trois, permettent de former des codons. Chacun d'entre eux sera ensuite traduit en acide aminé. La correspondance entre les codons et les acides aminés est donnée par le code génétique. La traduction est initiée par la présence d'un codon START (AUG) et s'arrête lors de la présence d'un codon STOP(UAA, UAG ou UGA).

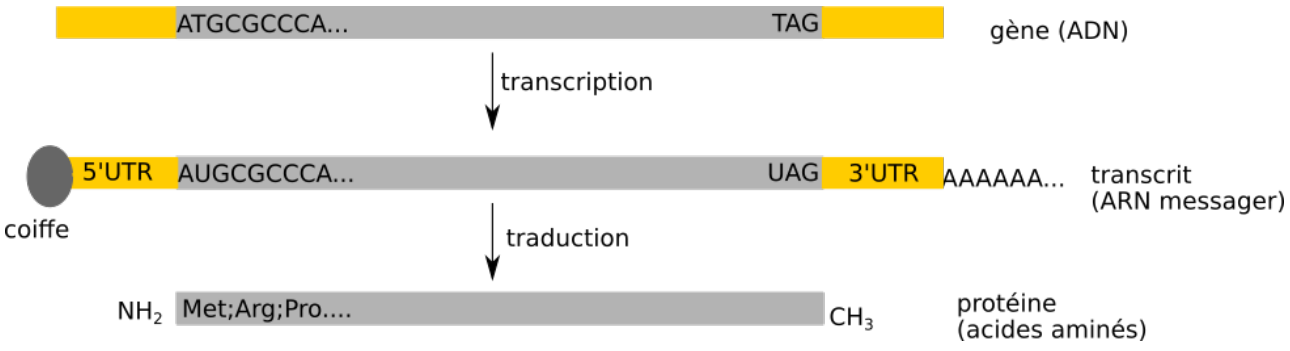


FIGURE 2 – Dogme central de la biologie moléculaire. Les gènes sont d'abord transcrit en pré-ARN messagers. Ceux-ci sont maturés puis traduit en protéines.

2 | Épissage

2.1 Définition générale

Chez les eucaryotes, les gènes sont structurés en exons et en introns. Les exons sont, la plupart du temps, codants et peuvent donc être traduits en protéines alors que les introns sont dits non codants. Lors de l'épissage, les introns sont retirés et les exons reliés entre eux grâce à l'intervention du spliceosome. La figure 3 présente l'exemple d'un gène composé de 4 exons et de trois introns. La molécule de pré-ARN messager contient encore tous les introns. Lors de l'épissage, ceux-ci sont excisés et les exons sont liés entre eux pour obtenir un transcrit mature. L'épissage à lieu dans le noyau des cellules eucaryotes. Il est souvent considéré comme co-transcriptionel [25] c'est-à-dire qu'il a lieu durant la transcription au fur et à mesure que le transcrit non mature est produit.

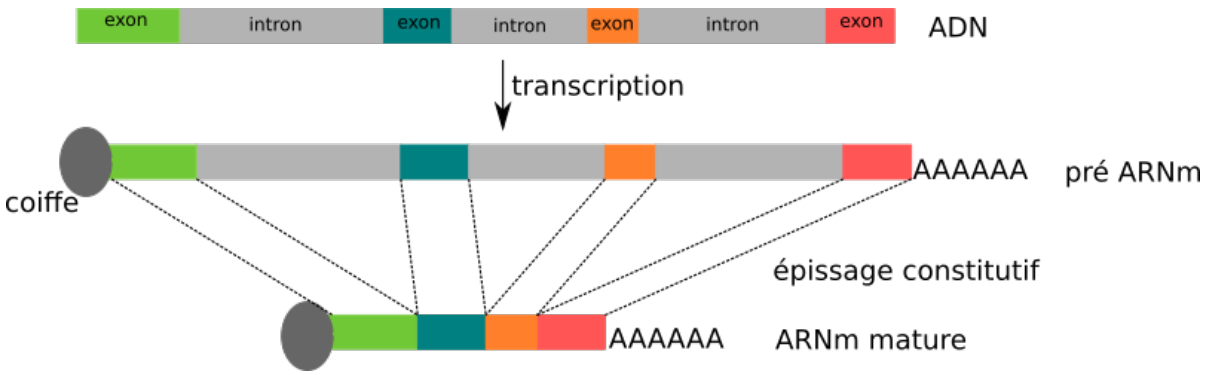


FIGURE 3 – Exemple d'un gène transcrit en pré-ARNm puis épissé.

2.2 Mécanismes moléculaires

2.2.1 Réaction d'épissage

La reconnaissance des sites d'épissage et l'excision des introns est réalisée par le spliceosome. Le spliceosome est un complexe protéique comprenant plusieurs centaines de protéines interagissant les unes avec les autres [24] ainsi que 5 ribonucléoprotéines (snRNP) [81]. Il existe plusieurs versions de ce complexe protéique, on s'intéresse ici au spliceosome majeur (nommé aussi spliceosome U2) qui reconnaît les sites d'épissage dits canoniques (GT et AG).

La réaction d'épissage est complexe, on peut la résumer en trois étapes principales (Figure 5). Dans un premier temps, les sous-unités U1 puis U2 du spliceosome s'assemblent autour de l'intron concerné, sur les sites d'épissage canoniques. Les sous-unités U4, U6 et U5 viennent ensuite s'assembler à leur tour sur l'intron. Ensuite, les sous-unités U1 et U4 se détachent du complexe protéique et suite à deux réactions de trans-esterification, l'excision de l'intron peut avoir lieu. Pour cela, le simple brin d'ARNm correspondant à l'intron est refermé sur lui-même sous forme de lasso. Les exons sont alors reliés entre eux à la suite de cette deuxième réaction.

La reconnaissance des bornes des introns se fait via celle des sites d'épissage mais aussi grâce à un autre motif, présent en amont du site accepteur d'épissage (c'est-à-dire à l'extrémité 3' de l'intron) : le site de branchement. Ce motif est notamment composé d'une Adénine très conservée. Le site de branchement est reconnu par la sous-unité U2 lorsqu'elle se fixe sur le brin de pré-ARNm. Le site d'épissage donneur (c'est à dire à l'extrémité 5' de l'intron) canonique est composé de la séquence GU et le site d'épissage accepteur de la séquence AG. Ces di-nucléotides font partie d'un motif consensus plus large. Plus celui-ci est conservé et plus les sites d'épissage sont dits forts. À l'inverse, plus il y a de divergences observées entre la séquence autour du site d'épissage et le motif consensus et plus le site est dit faible [83].

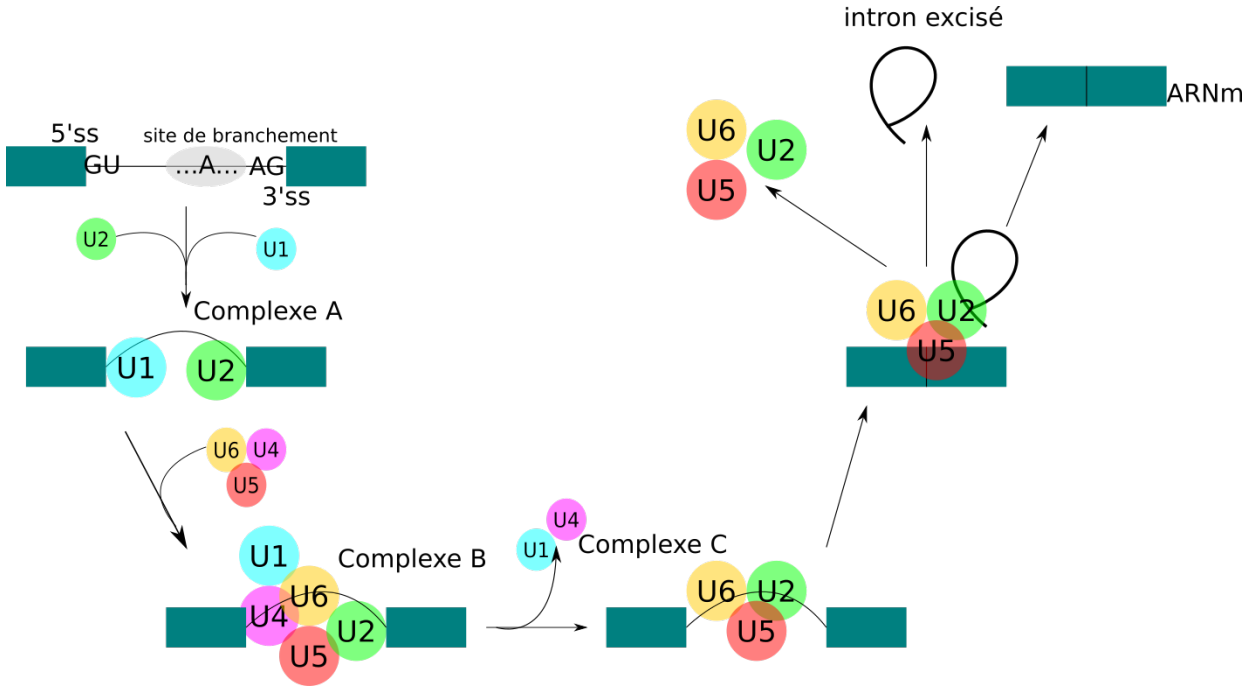


FIGURE 4 – Réaction d'épissage simplifiée. Seuls les snRNA sont représentés, mais le spliceosome est constitué aussi de nombreuses protéines. Figure adaptée depuis Wahl et al, 2009 [81]

.

Les sites d'épissage dits forts sont reconnus plus facilement et donc plus fréquemment que les sites d'épissage faibles.

2.2.2 Reconnaissance des exons

Dans certains génomes eucaryotes, comme chez l'Homme par exemple, les exons sont courts (de quelques dizaines à quelques centaines de paires de bases) et les introns peuvent être très longs (plusieurs milliers de paires de bases).

Il est alors courant que cela soit les exons qui soient reconnus plutôt que les introns ("exon définition")[62]. Dans ce cas, un complexe formé des snRNP U1 et U2 ainsi que plusieurs protéines s'assemblent autour de l'exon. Ce n'est que lors d'une deuxième étape, afin de relier les exons entre eux, que les snRNP U4, U5 et U6 sont recrutés [10].

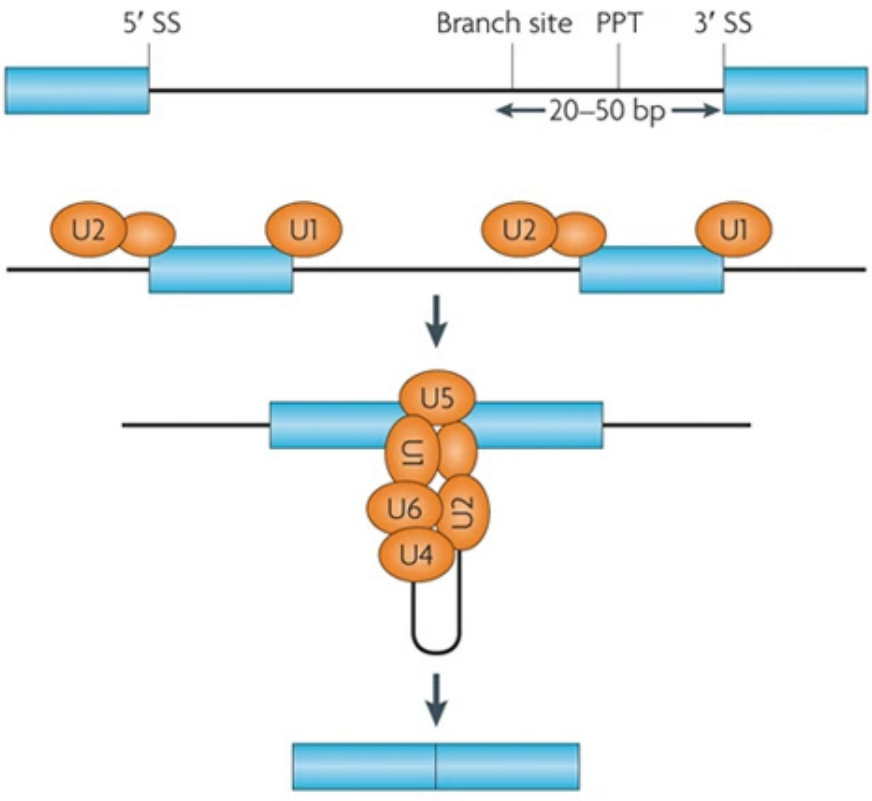


FIGURE 5 – Reconnaissance des exons. U1 se fixe sur le site donneur de l'exon alors que U2 se fixe sur le site accepteur. Un complexe se forme alors autour de l'exon. U4, U5 et U6 sont ensuite recrutés et le spliceosome s'assemble alors autour de l'intron. Figure adaptée depuis Keren et al, 2010 [31]

.

2.3 Épissage alternatif

2.3.1 Définition

D'après les définitions précédentes, on s'attend à obtenir un transcrit par gène, composé de chacun des exons du gène. Cependant, il arrive aussi que, pour un même gène, une ou plusieurs régions génomiques ne soient pas incluses dans tous les transcrits produits. Par exemple, un exon peut être inclus dans certains transcrits et exclu dans d'autres. Plusieurs

transcrits alternatifs sont alors produits depuis un même gène. On appelle cela l'épissage alternatif. Lorsque les bornes des exons sont conservées entre tous les transcrits d'un même gène, l'épissage est dit constitutif. La figure 6 reprend l'exemple du gène donné figure 3 mais illustre la possibilité de produire trois transcrits alternatifs depuis ce gène. Dans le premier cas, tous les exons sont conservés, dans le deuxième cas, l'exon bleu est exclu du transcrit présenté et dans un troisième exemple le deuxième intron est retenu. En effet, plusieurs sous-types d'évènements d'épissage alternatif existent.

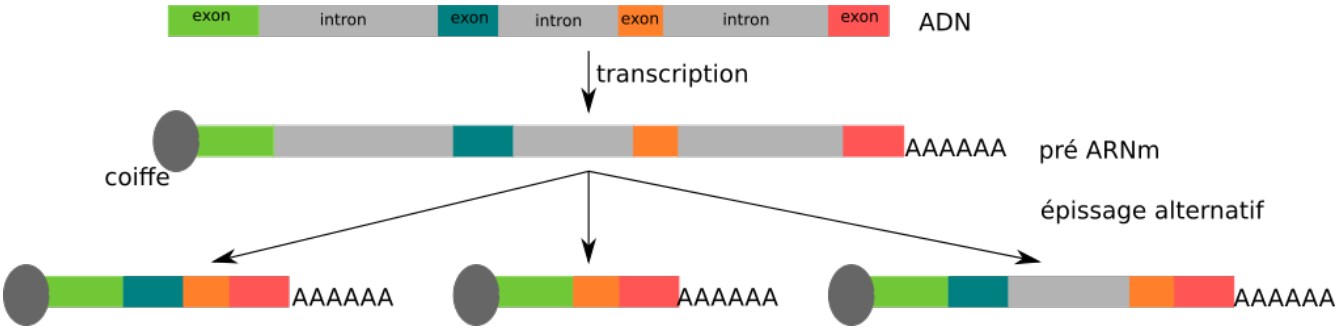


FIGURE 6 – Exemple d'un gène transcrit en pré-ARNm puis épissé alternativement. L'épissage alternatif permet de produire ici trois transcrits matures différents à partir d'un même gène.

2.3.2 Les différents types d'évènements d'épissage alternatif

Lorsque l'on compare, deux à deux, les transcrits alternatifs produits depuis un gène donné, on peut définir plusieurs sous-types d'événements d'épissage alternatif. Ils sont présentés figure 7. L'inclusion alternative d'un exon dans les transcrits est nommée saut d'exon. On parle de rétention d'intron lorqu'un intron peut être retenu. Les bornes des exons peuvent être modifiées par la présence de sites d'épissage donneur ou accepteur alternatifs. Certains exons ne sont jamais présents ensemble dans un même transcrit, ils sont alors dits mutuellement exclusifs. Il est enfin possible que deux exons ou plus soient exclus, l'un à la suite de l'autre, on parle alors de saut d'exon multiple.

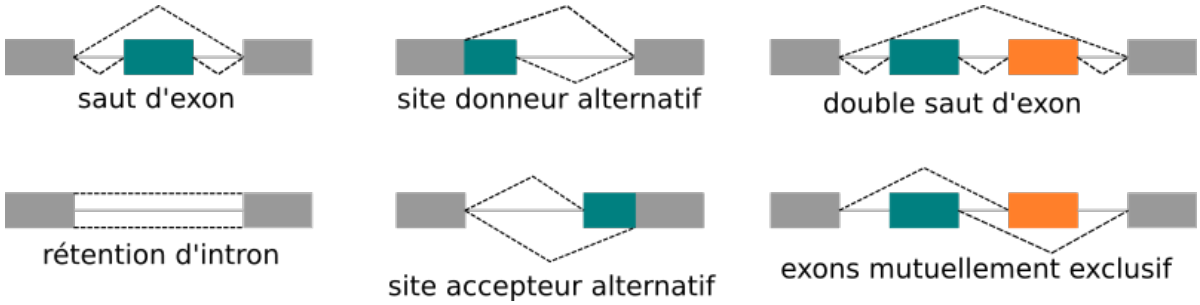


FIGURE 7 – Les différents types d'évènements d'épissage simples

Enfin, il peut arriver que, localement, plusieurs de ces sous-types d'évènements d'épissage soient combinés et ainsi qu'un seul évènement d'épissage permette la production d'au moins 3 transcrits. On observe alors des variations d'épissage dites complexes entre deux exons constitutifs d'un même gène.

2.3.3 Prévalence de l'épissage alternatif

On constate une prévalence importante de l'épissage alternatif dans les transcriptomes eucaryotes. Par exemple, [57] a permis d'estimer que 95% des gènes considérés comme multi-exoniques chez l'homme produisent au moins deux transcrits alternatifs. Les gènes produisant plusieurs transcrits alternatifs ne sont donc pas des cas isolés et ce résultat est retrouvé chez d'autres organismes. Même si les gènes des organismes invertébrés et des plantes semblent, en proportion, moins soumis à l'épissage alternatif (environ 20% chez *Arabidopsis Thaliana* ou chez *drosophila melanogaster*) une très grande diversité d'organismes eucaryotes est concernée [33].

2.4 Régulation de l'épissage alternatif

2.4.1 Erreur de la machinerie d'épissage et régulation

Tous les évènements d'épissage observés dans la cellule ne sont pas régulés par la machinerie cellulaire [59]. En effet, si certains évènements d'épissage alternatifs sont fonctionnels c'est-à-dire qu'ils permettent de produire plusieurs transcrits puis plusieurs protéines, d'autres ne sont dus qu'à des erreurs commises par la machinerie cellulaire. La quantification de la part d'évènements d'épissage alternatif régulés et celle de ceux dus au bruit de la machinerie d'épissage constituent actuellement une question toujours ouverte en biologie évolutive [77].

Un argument pour considérer un transcrit comme fonctionnel est qu'il permet de produire une protéine distincte de celle déjà produite par les autres transcrits du même gène. Cependant Tress et al [77] relèvent une faible proportion de gènes produisant plusieurs protéines (246 isoformes pour 12 716 gènes étudiés). Chez certaines espèces, comme *Vigna radiata* (le haricot mungo), l'épissage est même considéré comme principalement stochastique et très peu de gènes soumis à l'épissage alternatif semblent être régulés (environ 2.8% des gènes annotés) [68]. De plus, une autre étude permet de montrer qu'un grand nombre de gènes produit un transcrit principal [21]. Les autres transcrits produits sont moins abondants et peuvent donc être considérés comme mineurs.

Cependant, on connaît aussi de nombreux exemples de gènes dont les transcrits alternatifs permettent de produire des protéines différentes dont les fonctions sont distinctes. On peut par exemple citer le gène FAS, chez l'Homme, connu pour produire deux protéines : l'une soluble et pro-apoptotique (c'est-à-dire conduisant à la mort de la cellule) et l'autre transmembranaire et anti-apoptotique [27].

Deux points de vue co-existent actuellement autour de cette question. L'un affirmant que

de nombreux exemples sont clairement régulés, l'autre que les variations d'épissage sont, en grande majorité dues au hasard [69].

Une méthode récente (TRIFID) basée sur un algorithme de classification supervisée permet d'estimer la part des isoformes fonctionnels (c'est-à-dire traduits en protéines) ou non [60]. Cette méthode, basée sur l'analyse de données protéomiques, prend en compte la conservation des transcrits alternatifs entre espèce ainsi que la conservation des domaines protéiques dans les transcrits alternatifs. Enfin, identifier des évènements d'épissage dont les quantifications diffèrent significativement entre deux conditions expérimentales peut constituer une approche permettant d'identifier les événements d'épissage alternatifs qui sont régulés et donc fonctionnels dans le contexte des conditions étudiées (voir partie 3 de la thèse).

2.4.2 Épissage et NMD

Une partie des transcrits ne permettent pas d'aboutir à des protéines fonctionnelles. En effet, si une variation d'épissage induit l'inclusion ou l'exclusion d'une séquence dont la taille n'est pas un multiple de 3 alors son inclusion provoque un décalage du cadre de lecture sur toute la suite du transcrit. Il n'est pas rare que cela génère l'apparition d'un codon STOP précoce dans la séquence. Ainsi la protéine produite sera tronquée et donc, dans une très grande majorité des cas non fonctionnelle. Il est aussi possible qu'un codon STOP déjà présent dans la partie variable soit directement inclus au transcrit, sans décalage préalable du cadre de lecture. Ces transcrits sont pris en charge par la machinerie cellulaire et dégradé. Ce mécanisme est nommé NMD pour *Nonsense-mediated mRNA decay* [48].

L'identification des transcrits présentant un codon STOP précoce n'est pas simple. Chez les eucaryotes, avant l'exportation des transcrits matures dans le cytoplasme, des complexes protéiques, nommés EJC (pour *exon–exon junction complex*) sont ajoutés à chaque jonction d'exons [38]. Ensuite, la traduction a lieu en deux étapes. Une première étape ("Pionnering round of translation") permet de décrocher les EJC. La seconde de traduire les transcrits en protéines. Cependant, si un codon STOP est présent avant le dernier EJC, la première étape

sera stoppée avant de rencontrer cet EJC. Celui-ci restera donc en place sur le transcrit. La présence d'un EJC sur un transcrit lors de la deuxième passe de la transcription permet de recruter le NMD.

Ce mécanisme, en plus d'être connu pour dégrader les transcrits résultant d'erreurs de la machinerie d'épissage est aussi connu pour permettre la régulation de l'expression de gènes via la dégradation de certains transcrits fonctionnels. Le mécanisme RUST permet notamment d'expliquer la régulation de l'expression de nombreux facteurs d'épissage [37].

3 | Séquençage du transcriptome

Si le génome est le même dans toutes les cellules d'un individu, le transcriptome -c'est-à-dire l'ensemble des transcrits- diffère selon les tissus en fonction des gènes qui y sont exprimés. Séquencer le transcriptome plutôt que le génome permet donc d'avoir accès à ces différences. De plus, séquencer les transcrits permet aussi de les quantifier et donc de pouvoir décrire et comparer l'expression des gènes dans différents tissus. Enfin, séquencer les trancrits permet aussi de pouvoir les comparer entre eux et ainsi identifier les transcrits alternatifs pour chaque gène. Cependant, les techniques de séquençage évoluent et jusqu'à peu il n'était pas possible de séquencer des transcrits complets. Nous nous intéressons ici à l'évolution de ces techniques et aux caractéristiques des données obtenues avec chacune d'entre elles afin d'identifier les questions biologiques auxquelles elles peuvent aider à répondre.

Les techniques présentées ci-dessous permettent de séquencer, sur le même principe de fonctionnement, des molécules d'ADN (DNAseq) ou bien d'ARN (directement ou via la synthèse d'ADN complémentaire, RNAseq). On s'intéresse ici plus spécifiquement aux protocoles permettant de séquencer l'ARN.

3.1 Historique

Le premier séquençage d'une molécule d'ADN à été réalisé par Frederick Sanger en 1951. La mise en place, à plus grande échelle, de cette technique de séquençage à permis le premier séquençage d'un génome de bacteriophage en 1977 [67]. Dans les années 2000, l'avènement

des technologies de séquençage dites de nouvelle génération (NGS) à permis de réduire drastiquement les coûts du séquençage et ainsi de permettre la génération massive de séquences. L'accès aux séquences nucléotidiques que ce soit des génomes ou des transcriptome à permis des avancées importantes en biologie moléculaire et évolutive mais a aussi motivé de nouveaux développements méthodologiques. Aujourd'hui deux générations de technologies de séquençage co-existent : les technologies dites de deuxième (2GS) et de troisième génération (3GS).

3.2 Séquençage de 2eme génération

Plusieurs technologies de séquençage de deuxième génération co-existent (454, Ion Torrent, Illumina). Ces technologies permettent de séquencer l'ADN (DNAseq) comme l'ARN (RNAseq). On s'intéresse ici au fonctionnement de la technologie la plus couramment utilisée actuellement : Illumina.

3.2.1 Préparation des librairies

Différents protocoles existent en fonction du type d'ARN que l'on souhaite séquencer (ARN ribosomique, ARN total, transcrits de quelques gènes d'intérêt...). On se focalise ici sur le séquençage du transcriptome entier (c'est à dire des ARN messagers). Mis à part quelques exceptions, comme les gènes d'histones, ceux-ci sont reconnus dans les échantillons grâce à leur queue polyA. Tous les ARN dits polyA+ sont alors sélectionnés pour la suite des analyses. Cette étape permet aussi de ne pas sélectionner les ARN ribosomaux qui ne sont pas polyadénilés. Ils sont en effet très nombreux dans la cellule et s'ils sont séquencés une proportion importante de lectures proviendra des ARN ribosomiques dans les données. On dispose d'un nombre de lectures fini et, la plupart du temps, on souhaite plutôt séquencer des ARN messager.

Le protocole le plus utilisé pour la préparation de telles librairies est le protocole TruSeq. Ce protocole à été utilisé pour générer les données présentées dans la partie 2 de la thèse.

Nous le présentons ici.

L'ARN est une molécule bien moins stable que l'ADN. Ainsi, les simples brins d'ARNm sont, dans un premier temps, fragmentés puis rétrotranscrits en ADN complémentaire (double brin) grâce à une enzyme appelée transcriptase inverse. Des adaptateurs sont ensuite insérés à chaque extrémité des fragments obtenus. Ceux-ci permettent la réalisation des étapes suivantes du protocole mais aussi, par exemple, de reconnaître les molécules appartenant à un même échantillon (barcode). Les fragments sont ensuite fixés sur une lame de verre nommée flowcell. Enfin, une amplification PCR permet d'obtenir différents clusters de fragments identiques les uns au autres (Figure 8). Lors de cette amplification, les adaptateurs permettent à chacun des brins de se fixer, via leurs deux extrémités à la flowcell et ainsi de former un pont. La présence d'amorces, d'ADN polymérase ainsi que de nucléotides dans le milieu permettent alors la synthèse du brin complémentaire. Les doubles brins d'ADNc ainsi obtenus sont ensuite dénaturés et l'opération peut recommencer pour chacun des deux brins obtenus. L'enchaînement de plusieurs cycles d'amplification permet d'obtenir des clusters d'environ un millier de molécules d'ADN complémentaire identiques. Les fragments sont ainsi prêts à être séquencés.

3.2.2 Séquençage

Le séquençage illumina se fait via la synthèse des brins complémentaires à chacun des brins présents dans les clusters obtenus après l'amplification PCR. Un nucléotide est séquencé lors de chacun des cycles de séquençage. Lors de chacun de ces cycles, la présence d'une ADN polymérase ainsi que de nucléotides (dNTP) dans le milieu permet la synthèse du brin complémentaire. Celle-ci est réalisée grâce à des nucléotides modifiés. Ceux-ci portent un fluorochrome dont la couleur varie en fonction de la base nucléotidique. Ces nucléotides modifiés sont dits terminateurs réversibles. Cela permet qu'un seul nucléotide fluorescent soit ajouté par cycle. Une image est ensuite capturée et stockée à la fin de chaque cycle. L'étape d'amplification décrite plus haut, permet alors que le signal lumineux soit assez fort

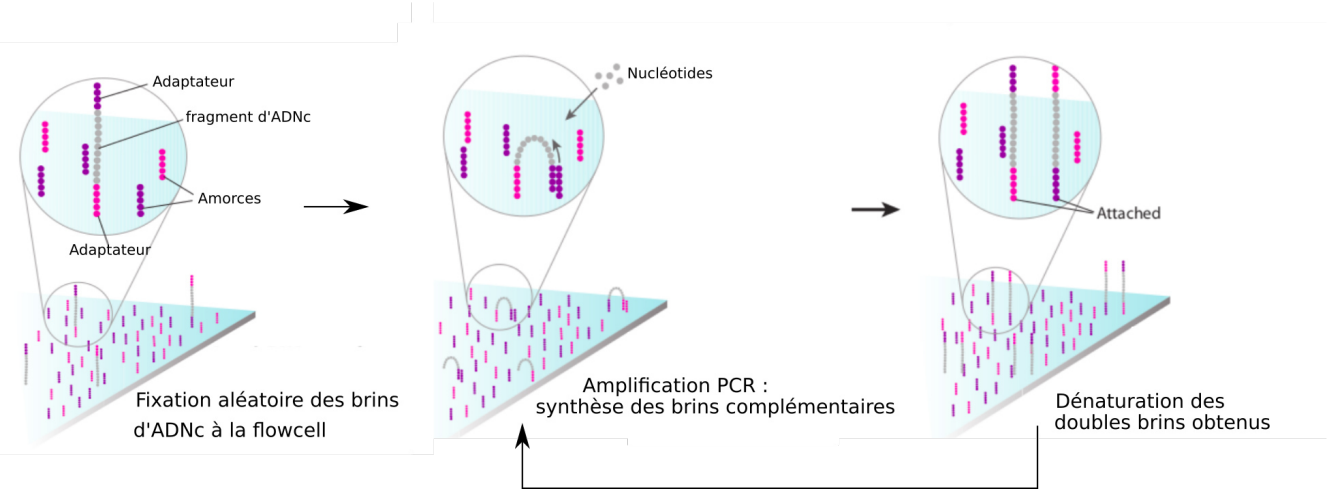


FIGURE 8 – Préparation de librairies illumina en vue du séquençage. figure adaptée depuis Mardis et al, 2008 [50]

pour être détecté. Le cycle suivant peut alors débuter. Le séquençage s'arrête au bout d'un nombre de cycles prédéfinis. Celui-ci correspondra à la longueur des lectures obtenues.

Les données brutes en sortie du séquenceur correspondent donc à une suite d'images capturées à la fin de chacun des cycles de séquençage. L'analyse de ces images permet de déterminer la nature des nucléotides insérés sur chaque cluster pour chaque cycle. Mises bout à bout ces images permettent de reconstituer la séquence des brins d'ADN séquencés (Figure 9).

3.2.3 Caractéristiques des données

Actuellement la technologie illumina est celle qui permet d'obtenir la profondeur de séquençage la plus importante (plusieurs centaines de millions de lectures par run de séquençage). Les lectures obtenues sont courtes (le plus souvent entre 75 et 250 paires de bases) mais celles-ci peuvent être "paired-end". Dans ce cas, les fragments d'ADNc sont séquencés depuis chaque extrémité. La longueur de l'insert (partie non séquencée entre les deux lectures) est alors variable. Enfin, les lectures obtenues sont de très bonne qualité puisque le

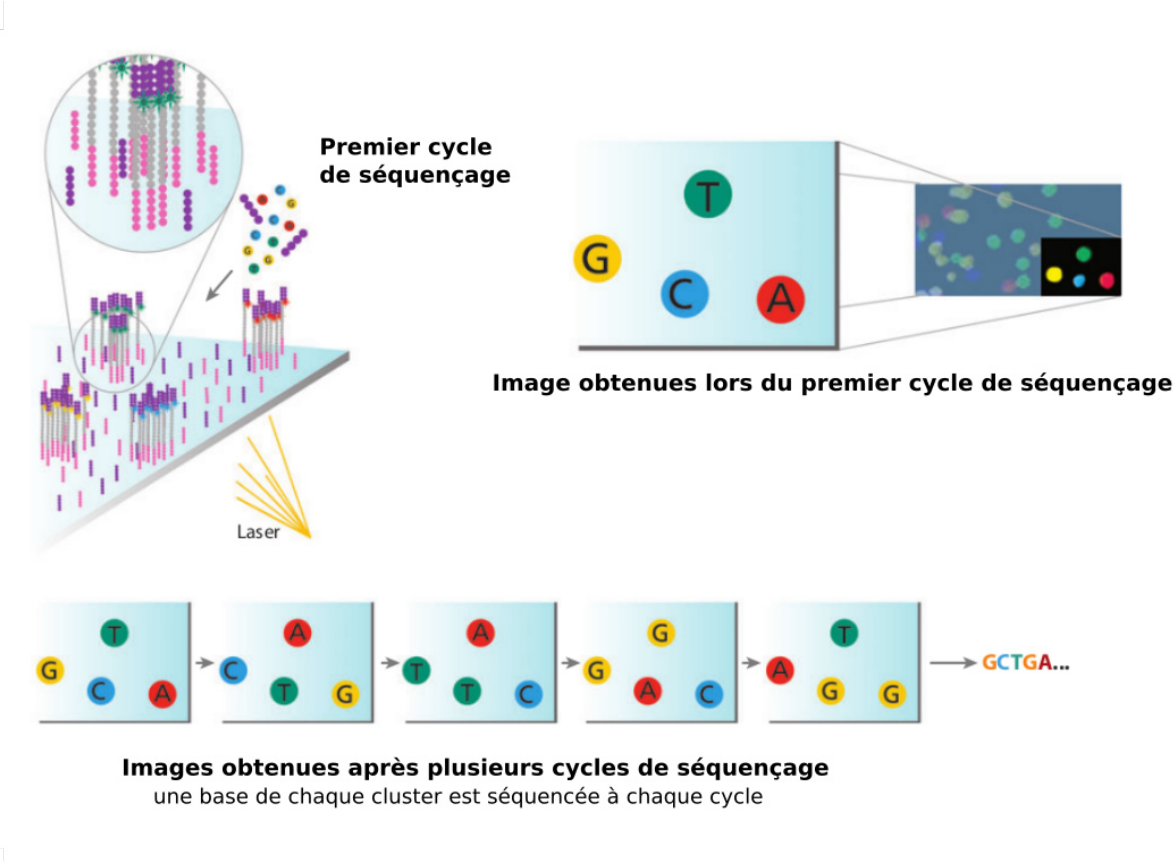


FIGURE 9 – Séquençage illumina. Figure adaptée depuis Mardis et al, 2008 [50]

taux d'erreur pour ce type de données est d'environ 0.5%.

3.3 Séquençage de 3eme génération

Les technologies de séquençage de troisième génération (3GS), apparues au début des années 2010, permettent d'obtenir des lectures longues (plusieurs milliers de paires de bases) et donc potentiellement, lorsque l'on fait du RNAseq, des transcrits complets. Deux technologies coexistent actuellement. Celle développée par l'entreprise Pacific Biosciences (PacBio) et celle développée par Oxford Nanopore Technologies (ONT). Les données présentées dans cette thèse (Partie 2) ont été obtenues avec la technologie Nanopore. Après une brève présentation de la technologie PacBio on se focalisera donc sur celle de Nanopore et plus

particulièrement sur les protocoles utilisés lors de la génération des données présentées dans la partie 2.

3.3.1 PacBio

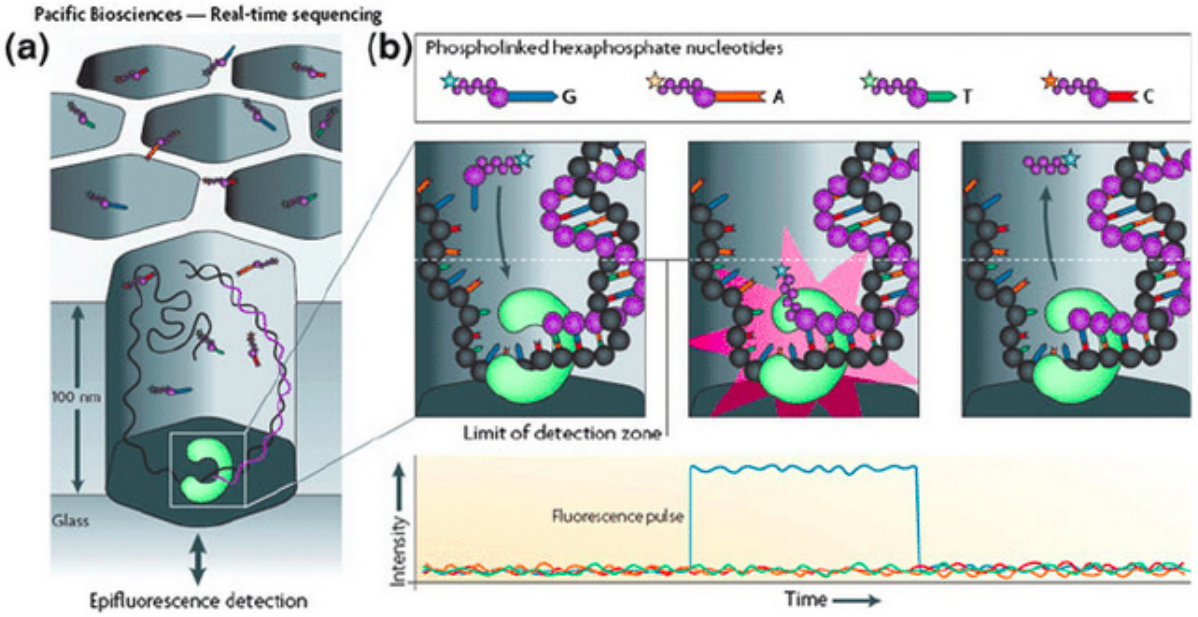


FIGURE 10 – Séquençage par Pacbio. La lecture du signal fluorescent associé à chaque nucléotide est faite en temps réel au fur et à mesure que le deuxième brin d'ADNc est synthétisé. Figure reproduite depuis Metzker et al, 2010 [54]

Tout comme Illumina, cette technologie est basée sur la synthèse du brin complémentaire à celui qui est séquencé (Figure 10). Cependant, cette technologie ne nécessite pas d'amplification PCR. Chacun des brins d'ADN est pris en charge par une seule ARN polymérase fixée au fond de chaque puits de la flowcell [54]. Les nucléotides intégrés un par un au brin d'ADN complémentaire portent des fluorochromes (une couleur distincte par base). C'est le brin d'ADN qui se déplace pour passer dans l'ADN polymérase et les nucléotides fluorescents sont ajoutés au fur et à mesure, sans pause entre chaque base. C'est pour cela que cette technologie est dite de séquençage en temps réel. L'ajout de chaque base fluorescente

permet l'émission d'un signal réceptionné par un capteur placé dans le puits. Ce signal peut ensuite être traduit en séquence. La figure 10 montre l'exemple de l'ajout d'une guanine au brin synthétisé. Un signal fluorescent dont l'intensité correspond à la couleur associée à la guanine permet de l'identifier. Ensuite, le fluorochrome se détache et la base suivante peut être intégrée au brin d'ADN synthétisé.

Un protocole, nommé Iso-Seq (pour "Isoform Sequencing") [20] permet d'adapter la technologie de séquençage pour le RNAseq. Comme dans les cas d'Illumina, le séquençage de l'ARN se fait après une étape de rétrotranscription de l'ARN en ADN complémentaire. Ce protocole permet notamment de circulariser l'ADNc et de séquencer plusieurs fois à la suite la même molécule. Cela permet de réduire le taux d'erreur obtenu en fin de séquençage.

3.3.2 Oxford Nanopore

La seconde technologie permettant de produire des lectures longues est celle développée par Oxford Nanopore Technologies. Comme celle présentée ci-dessus, cette technologie permet le séquençage de l'ARN après une étape de rétrotranscription (protocole cDNA) mais elle permet aussi de séquencer les simples brins d'ARN directement (protocole directRNA) [18]. Les protocoles de préparation des librairies dans ces deux cas sont présentés ci-dessous. On s'intéressera ensuite au fonctionnement des séquenceurs puis aux caractéristiques des données obtenues.

Préparation des librairies cDNA

Le protocole le plus classique proposé par Nanopore est celui proposé pour séquencer de l'ADN complémentaire (ADNc). Celui-ci, comme pour les données de 2e génération, implique une première étape de rétrotranscription des simples brins d'ARN messagers. Contrairement au protocole illumina, il n'y a pas ici d'étape de fragmentation. Cela permet de séquencer des lectures longues et donc potentiellement des transcrits complets. Ce protocole est présenté figure 11. Dans un premier temps, des adaptateurs sont hybridés aux simples brins d'ARNm.

Ceux-ci permettent d'initier l'étape de rétrotranscription puis la synthèse du brin complémentaire et l'amplification des transcrits par PCR. On peut noter qu'une version plus récente de ce protocole existe. Elle permet de préparer les librairies sans étape d'amplification PCR afin d'éliminer les biais induits par cette technique. C'est bien la version présentée figure 11 qui a été utilisée dans le cadre de la production des données présentées en partie 2 de cette thèse. Enfin des adaptateurs de séquençage sont ajoutés aux extrémités des doubles brins d'ADNc obtenus lors de l'étape précédente. Ces molécules d'ADNc sont alors prêtes à être séquencées. Des protocoles, plus récents que ceux utilisés pour générer les données présentées dans la partie 2 de la thèse, permettent de préparer des librairies brin-spécifiques.

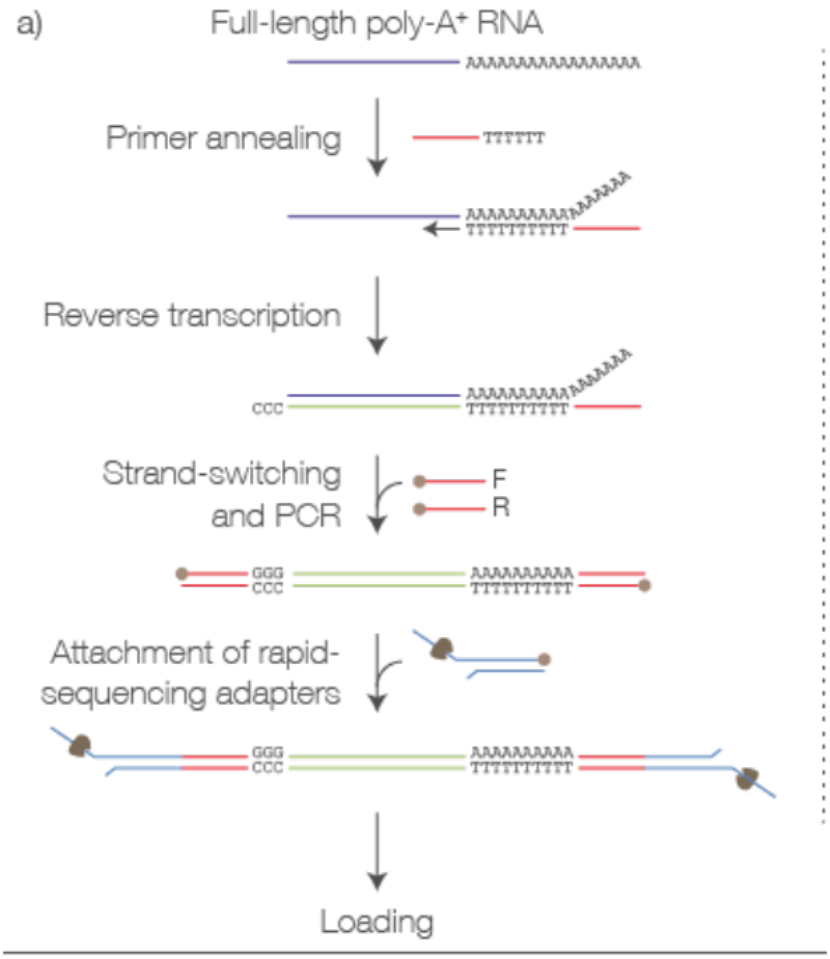


FIGURE 11 – Préparation des librairies nanopore cDNA source : nanoporetech.com

Préparation des librairies RNA direct

L'un des intérêts principaux de la technologie Nanopore est de pouvoir séquencer des brins d'ARN messager directement, sans passer par une étape de rétrotranscription (RT) en amont. Cela permet d'éviter de potentiels biais induits par l'étape de RT. Cette technologie récente évolue rapidement et même si une quantité importante d'ARN (500 ng d'ARN polyA+ alors qu'il suffit d'1 nanogrammes d'ARN polyA+ pour le protocole cDNA) est nécessaire au départ pour confectionner les librairies, les profondeurs de séquençage obtenues évoluent et augmentent rapidement. Cette technique peut aussi permettre d'accéder aux modifications épigénétiques des molécules d'ARN séquencées. Le protocole de préparation de ce type de librairie est présenté figure 12. Il est à noter que les brins d'ARNm sont tout de même soumis une étape de reverse transcription (notée comme optionnelle sur la figure). Celle-ci permet de stabiliser le simple brin d'ARN en synthétisant le brin d'ADN complémentaire. Cependant, c'est bien le brin d'ARN natif qui passe dans le pore et qui sera séquencé. Ensuite, des adaptateurs de séquençage sont ajoutés à l'extrémité des doubles brins ainsi générés. L'ARN peut alors être séquencé.

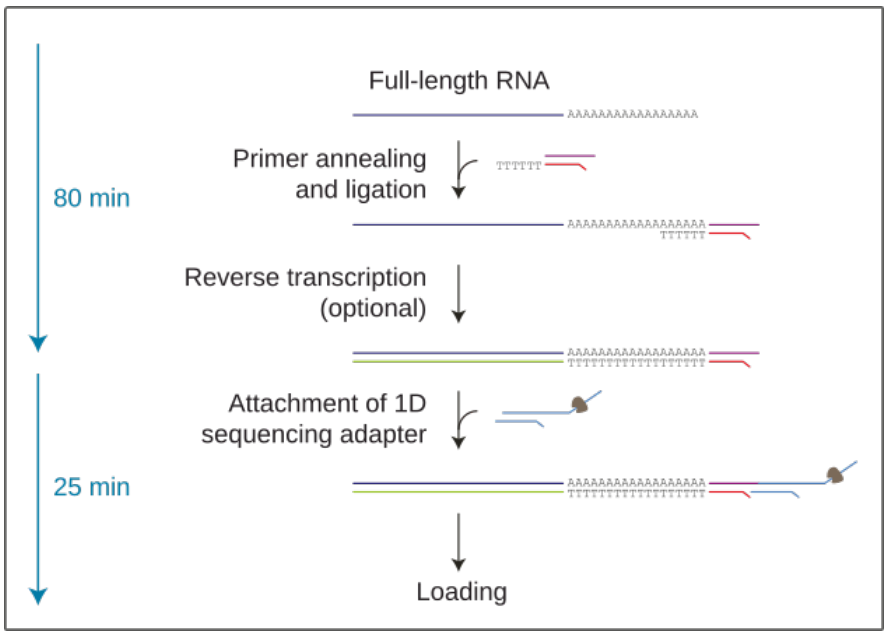


FIGURE 12 – Préparation de librairies nanopore RNAdirect source : nanoporetech.com

Teloprime

Nous nous intéressons ici brièvement au protocole dit Teloprime. Celui-ci a été utilisé pour générer un des jeux de données présenté dans la partie 2. Ce protocole permet d'enrichir les librairies en transcrits complets c'est-à-dire coiffés et polyadénylés. En effet, toutes les molécules d'ARN extraites des cellules étudiées ne sont pas nécessairement complètes. Celles-ci peuvent être en partie dégradées dans la cellule. Enrichir les librairies en transcrits complets permet de pouvoir espérer obtenir un pourcentage de lectures correspondant à des transcrits complets plus important mais cela induit aussi d'autres biais (voir l'article présenté partie 2).

Séquençage

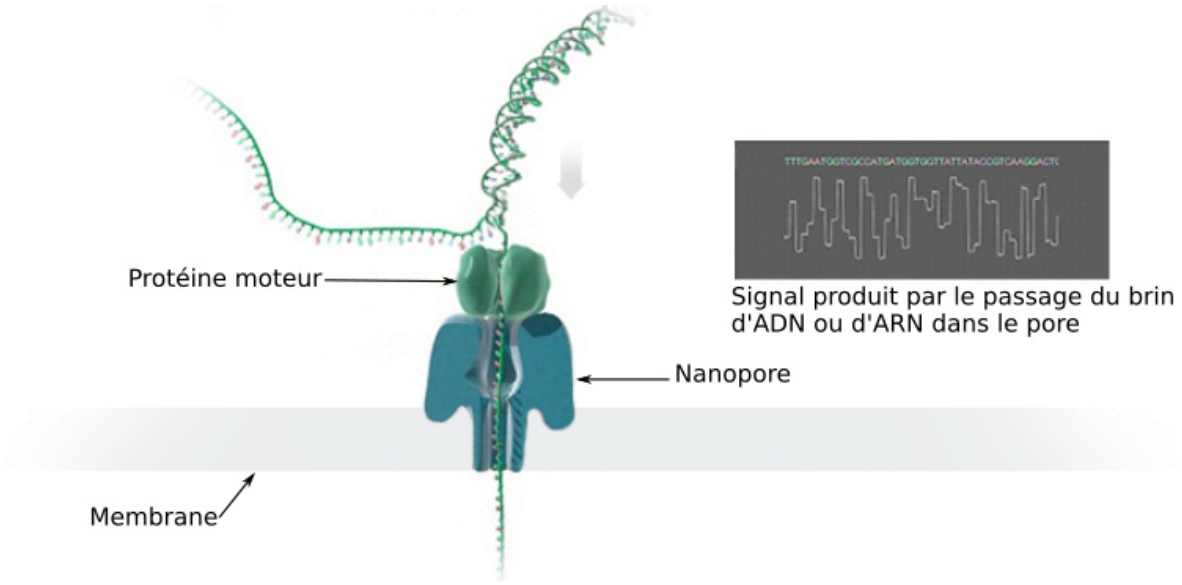


FIGURE 13 – Séquençage Nanopore. Le brin d'ARN ou d'ADNc est guidé dans le pore par une protéine moteur. Le signal produit par le passage de la molécule dans le pore est représenté dans le cadre gris. Une étape de basecalling permet d'interpréter ce signal et d'en déduire la séquence du brin séquencé. Figure adaptée depuis nanoporetech.com

On s'intéresse maintenant à la façon dont les librairies sont séquencées. Plusieurs séquen-

ceurs existent mais ils sont tous basés sur le même principe : Le passage d'un brin d'ADNc ou d'ARN dans un pore biologique [73]. Le pore est un complexe protéique permettant dans un premier temps de déplier le brin d'ADN ou d'ARN à séquencer puis d'identifier la base nucléotidique présente dans la deuxième protéine composant le pore. Celle-ci contient un adaptateur permettant à la base nucléotidique qui est séquencée de rester en place assez longtemps dans le pore pour être identifiée. Un flux constant d'ions circule à l'intérieur du pore et le passage de chacune des bases nucléotidiques provoque une altération du courant électrique spécifique [61]. En fonction du courant mesuré dans le pore, on peut déduire la nature de la base nucléotidique qui y passe. Une étape dite de base calling, permet d'interpréter le signal obtenu et ainsi de reconstruire la séquence du brin considéré.

Caractéristiques des données obtenues

Cette technologie permet de produire des lectures longues (jusqu'à plusieurs milliers de paires de bases) qui couvrent parfois des transcrits complets. Les principales limitations de cette technologie de séquençage de troisième génération sont le taux d'erreur (entre 8% et 12% actuellement) et la faible profondeur de séquençage. Cependant ces technologies évoluent très rapidement et les taux d'erreur baissent (ils étaient d'environ 15% au début de ce travail de thèse). La profondeur évolue aussi et augmente considérablement ces dernière années (voir les perspectives de la partie 2). Il est important de noter que les erreurs de séquençage ne sont pas aléatoires. On remarque notamment des erreurs au niveau des homopolymères, c'est-à-dire, lorsqu'une même base nucléotidique se répète plusieurs fois dans un transcrit. Dans ce cas-là, on observe souvent des erreurs sur le nombre de fois où la base est répétée.

4 | Analyse de l'épissage à partir des données RNAseq

L'analyse des données issues du séquençage du transcriptome (RNAseq) permet d'étudier les variations d'épissage dans les gènomes eucaryotes. Ces données permettent aussi de répondre à d'autres questions biologiques. Elles permettent par exemple de réaliser des analyses d'expression de gènes. En effet, quantifier les transcrits produits pour chacun des gènes du génome dans un tissu donné permet d'estimer le niveau d'expression des gènes. Les données RNAseq permettent aussi de détecter les SNPs (single nucleotide polymorphism : mutation d'une seule base dans le génome) ou bien les fusions de gènes dans le cas de tissus tumoraux par exemple. On se focalise ici sur l'étude de l'épissage alternatif.

4.1 Analyse des données de 2ème génération

Depuis les années 2000, les lectures courtes sont utilisées pour détecter et quantifier les variations d'épissage. De nombreuses méthodes ont été développées dans ce but. Celles-ci peuvent être basées sur un génome de référence ou non. On distingue aussi deux types de méthodes en fonction de l'échelle d'étude utilisée. Les méthodes peuvent être locales ou globales. On présente ici le fonctionnement de ces méthodes, leurs points communs et leurs différences.

4.1.1 Méthodes basées sur l'alignement

Une première approche afin de détecter les événements d'épissage alternatif dans les données RNAseq est d'aligner les lectures courtes sur un génome de référence. Les données RNAseq présentent la particularité de ne pas contenir les introns. Ainsi pour les aligner contre un génome de référence il est nécessaire d'autoriser l'ouverture de longs gaps dans les lectures contenant des jonctions d'exons (Figure 14). Pour définir les bornes des exons, on peut s'appuyer sur les jonctions déjà connues (c'est-à-dire les bornes des exons présents dans les annotations) mais aussi sur les sites d'épissage canoniques (GT, AG). Cependant ces dinucléotides peuvent aussi être présents dans le génome par hasard. La définition des bornes des exons n'est donc pas toujours facile.

FIGURE 14 – Alignement des lectures RNAseq. Les lectures chevauchant plusieurs exons sont alignées après l'ouverture d'un long gap. Ceux-ci sont représentés en pointillé. Les lectures exoniques sont quant à elles alignées en un seul bloc.

Des aligneurs adaptés ont été développés dans ce sens. On peut les classer dans deux catégories : les aligneurs dits *exon first* et ceux basés sur l'extension de graines (*seed and extend*). Les premiers alignent déjà un maximum de lectures sur les exons (dont les coordonnées sont préalablement définies dans les annotations). Ils alignent ensuite les lectures non alignées lors de la première étape sur les jonctions d'exons. Les aligneurs dits seed-and-extend alignent d'abord des graines, c'est à dire une sous-séquence des lectures, avant d'essayer de les rejoindre entre elles et ainsi d'étendre l'alignement autour des graines afin d'aligner la lecture complète. Par exemple, TopHat [75], utilisé sans les annotations, est un aligneur dit exon-first alors que HISAT [32] et STAR [15]sont basés sur des algorithmes de

seed and extend.

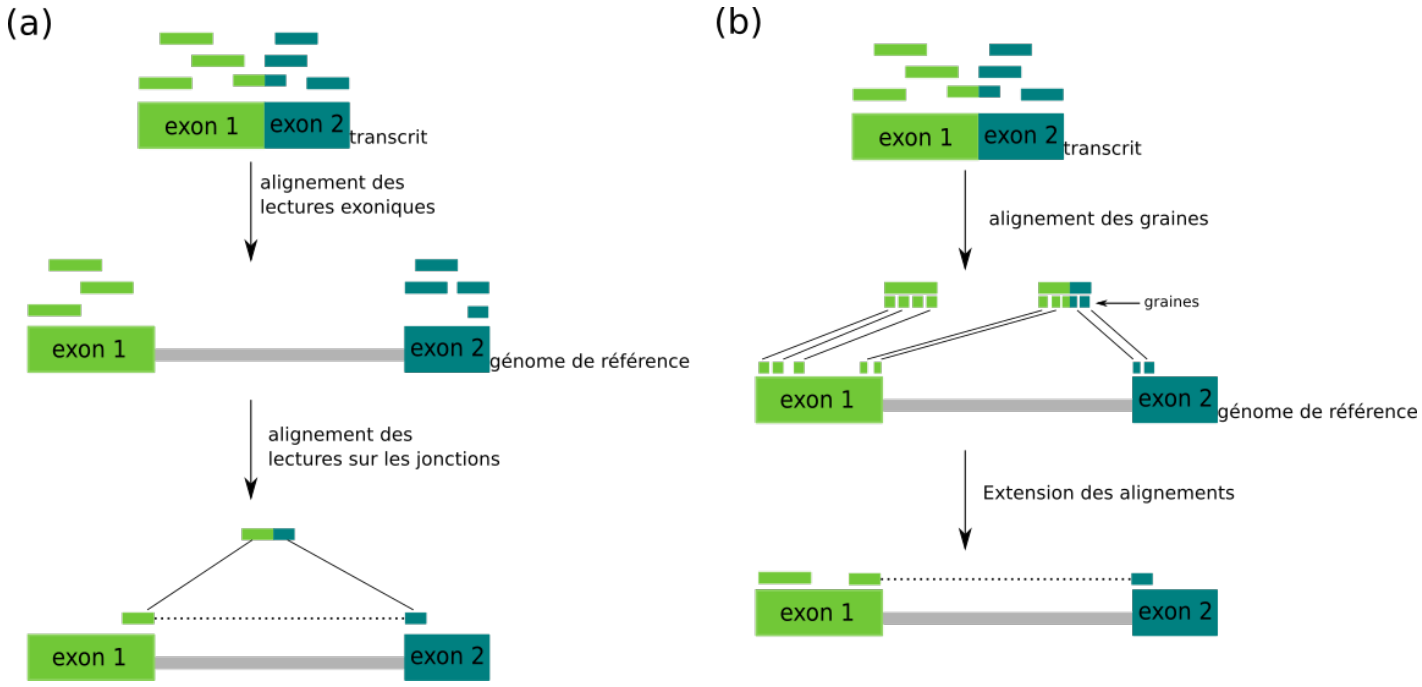


FIGURE 15 – Aligneurs *exon first* et *seed and extend*. (a) Les aligneurs dits *exon first* alignent d'abord les lectures exoniques puis se servent des premiers alignements pour aligner les lectures restantes sur les jonctions. (b) Les aligneurs dits *seed and extend* alignent déjà des graines, puis étendent les alignements autour des graines alignées. Figure adaptée depuis Grabber et al, 2011 [19]

Afin de prendre en compte les jonctions qui sont annotées mais aussi de nouvelles jonctions qui peuvent être présentes dans les données et ainsi mieux définir les bornes des exons, certains aligneurs comme STAR [15] proposent de réaliser les alignements en deux fois (STAR 2PASS). Une première passe permet d'aligner les lectures comme présenté ci-dessus en tenant compte des annotations. Une deuxième passe permet de réaligner les lectures en tenant compte cette fois-ci des nouvelles jonctions trouvées lors de la première passe, en plus de celles présentes dans les annotations. En considérant ces nouvelles jonctions *a priori* les bornes des exons sont mieux définies.

Une des limites de ces méthodes concerne les cas où il est impossible de choisir de quelle région génomique provient une lecture. On appelle cela l'alignement multiple (mapping multiple). En effet, une lecture peut être alignée à plusieurs endroits dans le génome. Chacun des alignements effectué est associé à un score prenant en compte le nombre de mismatches, la taille des insertions/délétions qu'il a fallu considérer afin de réaliser l'alignement. Parfois, un des alignements présente clairement un score plus élevé que les autres mais ce n'est pas toujours le cas. Il n'est en effet pas rare de retrouver des répétitions exactes dans les génomes. Dans ce cas, une même lecture peut s'aligner à plusieurs endroits différents et les scores associés à chacun de ces alignements seront les mêmes. Il est important de noter que d'un point de vue moléculaire, une lecture ne peut provenir que d'une seule région génomique. La plupart des aligneurs attribuent alors la lecture au hasard sur l'une des régions génomiques sur laquelle elle s'aligne. Cela revient à attribuer 50% des lectures à une position et 50% des lectures à l'autre position lorsque qu'il y a deux alignements ex-aequo par exemple. Cette façon de faire est héritée de ce qui est fait habituellement en DNAseq où l'on considère qu'il n'y a pas d'hétérogénéité de couverture. Cependant, dans les données RNAseq, la couverture dépend du niveau d'expression de chacun des transcrits. Attribuer aléatoirement une lecture revient donc à faire l'hypothèse qu'il n'y a pas de différence d'expression entre les différentes régions sur lesquelles la lecture s'aligne. Le problème du mapping multiple est important à considérer lors d'analyse de l'épissage : la qualité des alignements détermine la qualité des quantifications obtenues en aval.

La limite principale des aligneurs *exon first* est la sur-estimation du nombre de lectures s'alignant sur les pseudogènes rétrotranscrits [19]. Ces pseudogènes proviennent de la réinsertion dans le génome d'un transcrit préalablement rétrotranscrit en ADN. Ils ne possèdent donc pas d'intron et accumulent des divergences avec leur gène parent au fil du temps. Les problèmes de mapping multiple entre un pseudogène rétrotranscrit et son gène parent sont fréquents. En effet, lors de la première étape d'alignement, les lectures provenant des jonctions

d'exons du gène parent pourront s'aligner en seul bloc sur le pseudogène rétrotranscrit. On ne testera donc pas de les aligner sur une jonction d'exon lors de la deuxième étape et on ne les alignera donc jamais sur le gène parent.

4.1.2 Méthodes basées sur l'assemblage

Une deuxième approche consiste à assembler les lectures plutôt que de les aligner. Cela permet l'analyse de données provenant d'espèces non modèles mais aussi de trouver des évènements d'épissage non annotés chez les espèces modèles.

Pour cela, on liste tous les mots de taille k, nommés k-mers, présents dans les lectures. Les k-mers sont ensuite utilisés afin de construire un graphe de De Bruijn. Chacun des k-mers constitue un noeud et deux k-mers sont reliés entre eux par une arête s'ils se chevauchent (c'est-à-dire si le suffixe de l'un est identique au préfixe de l'autre) sur (k-1) nucléotides. Cette structure de données permet ensuite de reconstruire des transcrits complets ou bien de détecter des évènements d'épissage locaux grâce à différents algorithmes. Les méthodes permettant de détecter des événements d'épissage exon-centrés dans le graphe de De Bruijn sont détaillées dans la partie 3 (Modélisation des évènements d'épissage complexes).

Les méthodes basées sur l'assemblage sont complémentaires de celles basées sur l'alignement des lectures. [8] a permis de montrer que 70% des évènements de saut d'exon sont retrouvés par les deux types de méthodes. Les méthodes basées sur l'assemblage permettent de trouver de nouveaux exons et de nouveaux sites d'épissage alors que les méthodes basées sur l'alignement permettent de trouver plus de variants rares ainsi que plus d'évènements dans les régions répétées.

4.1.3 Échelle d'étude

Qu'elles soient basées ou non sur un génome de référence, les méthodes permettant d'étudier l'épissage alternatif peuvent être locales ou globales.

Les méthodes locales étudient des évènements d'épissage alternatif à l'échelle de l'exon. Pour cela, ces méthodes considèrent deux exons flanquants, entre lesquels elles étudient

les variations d'épissage. Parmi elles, on peut citer MISO [29], DEXseq [3] ou encore AS-Quant[16]. Ces dernières sont toutes basées sur un génome de référence. KisSplice [64] permet d'étudier des événements d'épissage locaux sans génome de référence. D'autres méthodes, plus récentes et exons centrées, toutes basées sur l'utilisation d'un génome de référence, permettent de considérer des événements d'épissage alternatif qui aboutissent à plus de deux transcrits. Ces évènements sont dits complexes. Il s'agit de MAJIQ[79], Whippet [72], Leafcutter[40], FRASER [53] ou encore McSplicer[2].

Les méthodes globales permettent de reconstruire des transcrits complets. Cette tâche est difficile avec des lectures courtes et ces méthodes commettent encore beaucoup d'erreurs [71]. Parmi les méthodes globales, on peut citer Stringtie [58], Cufflinks[76], FlipFlop [11], Scripture [23] qui sont toutes basées sur l'alignement des lectures sur un génome de référence. Oases [70], Trinity [22], Trans-ABySS [63] et rnaSPAdes [13] sont elles aussi des méthodes globales mais ne nécessitent pas de génome de référence.

Tous ces outils présentent des performances (temps de calcul, CPU et mémoire RAM utilisés) différentes mais aussi des taux de précision et de sensibilité divers. Les méthodes exon-centrées sont connues pour être plus précises que les méthodes globales [52]. Les assembleurs sont connus pour demander des ressources RAM importantes mais cela peut aussi être le cas de certains aligneurs. STAR a par exemple besoin de 30G de RAM pour reconstruire l'index du génome humain ainsi que pour aligner les lectures.

4.1.4 Identification des événements d'épissage/des transcrits

D'une part, dans le cas des méthodes locales, la détection des évènements d'épissage peut être faite grâce aux annotations dans le cas où les lectures ont préalablement été alignées. C'est ce qui est fait par les méthodes locales et basées sur un génome de référence citées dans le paragraphe précèdent. Les évènements sont ensuite catégorisés en différents types (saut d'exon, rétention d'intron etc. voir paragraphe 2.3.2). Dans le cas de KisSplice [64], méthode locale basée sur l'assemblage, les évènements d'épissage alternatif correspondent à des structures particulières dans le graphe de De Bruijn, nommées bulles. Ces bulles sont énumérées

par KisSplice. Chacun des deux chemins de la bulle (noté le chemin du haut de la bulle et le chemin du bas de la bulle) correspond à une séquence. Chaque bulle correspond donc à une paire de séquences qui partagent des régions flanquantes. L'un des chemins contient la séquence alternativement épissée et l'autre ne la contient pas. Dans le cas où un génome de référence existe, chacun des chemins peut être réaligné sur le génome de référence. KisSplice2refGenome, post-traitement de KisSplice, permet d'analyser les alignements obtenus afin de catégoriser les différents évènements détectés en amont.

D'autre part, lorsque l'on veut reconstruire les transcrits complets, dans le cas où l'on a accès à un génome de référence et où les lectures ont été alignées, un graphe de chevauchement peut être construit. Celui-ci permet de relier entre eux chacune des lectures alignées dont les positions génomiques se chevauchent. C'est ce que fait Cufflinks [76] par exemple. Chacun des transcrits correspond à un chemin du graphe mais tous les chemins ne correspondent pas nécessairement à des transcrits. Ainsi il n'est pas facile, avec des lectures courtes, de phaser les exons entre eux, c'est-à-dire d'identifier quels exons sont inclus ou exclus ensemble dans les transcrits.

Une autre structure de données permet de reconstruire les transcrits complets depuis l'alignement des lectures : le graphe d'épissage. C'est ce que reconstruisent Scripture [23] et StringTie [58] par exemple. Dans ce type de graphe, les noeuds représentent des exons (ou des partie d'exons dans les cas de site d'épissage donneur ou accepteur alternatif). Les arêtes représentent les jonctions possibles entre chacun des exons, c'est-à-dire les introns. Plus il y a d'arêtes dans une région donnée du graphe d'épissage et plus les variations d'épissage sont nombreuses.

Il est intéressant de noter que StringTie infère à la fois les transcrits et leur quantification alors que Cufflinks infère d'abord les transcrits, puis leur quantification.

Enfin les transcrits complets peuvent être reconstruits sans génome de référence. Cela peut être fait avec Oases [70] ou Trinity [22]. Ces méthodes sont toutes deux basées sur la reconstruction du graphe de De Bruijn. La qualité de l'assemblage obtenue dépend no-

tamment du choix de la taille des k-mers, c'est-à-dire de la valeur de k. En effet, si des répétitions de taille inférieure à k sont présentes dans les transcrits à assembler alors il sera difficile d'identifier leur localisation dans les transcrits finaux (cycles dans le graphe de De Bruijn). Aussi, une valeur de k trop élevée ne permettra pas de reconstruire des transcrits complets dans le cas ou les transcrits sont peu couverts. Ainsi, certains assembleurs comme Oases permettent d'assembler les transcrits successivement avec plusieurs valeurs de k et de construire des consensus par la suite.

4.1.5 Analyse différentielle

La plupart du temps, une fois les variations d'épissage identifiées, on souhaite détecter celles d'entre elles qui sont régulées. On cherche alors à identifier les évènements ou les transcrits différentiellement épissés entre les conditions expérimentales étudiées. D'un point de vue statistique, on pose l'hypothèse nulle suivante : Les transcrits alternatifs ont le même niveau d'expression entre les deux conditions.

Pour cela, il est nécessaire de quantifier chacun des évènements d'épissage ou des transcrits. Le plus souvent, cette étape de quantification est couplée avec celle de l'identification des évènements.

Quantification

Dans le cas des méthodes basées sur un génome de référence, on peut compter combien de lectures ont été alignées sur chacun des variants étudiés. Dans le cas des méthodes basées sur l'assemblage, il est possible de réaligner les lectures sur le graphe préalablement reconstruit et ainsi de quantifier le nombre de lectures soutenant chacun des noeuds ou des chemins considérés. Une fois les variants quantifiés, on souhaite comparer les quantifications obtenues dans chacune des conditions étudiées. Pour cela, il est nécessaire de disposer de plusieurs réplicats par condition afin de savoir si les variations observées entre conditions sont simplement dues à la variabilité technique ou biologique ou bien si elles sont effectivement régulées. C'est donc un point important à prendre en compte, en amont des analyses

bio-informatiques mais aussi en amont des manipulations expérimentales, lors de la planification expérimentale. Les quantifications sont ensuite normalisées. Cela permet de prendre en compte les variations de profondeur de séquençage entre les réplicats. La méthode pour normaliser les comptages la plus couramment utilisée est celle implémentée dans le paquet R DeSeq2 [46].

Détection des transcrits différentiellement épissés

On souhaite ici savoir si les variations d'épissage observées sont significativement différentes entre deux conditions expérimentales. Pour cela, différentes méthodes existent. On peut citer, par exemple, Cuffdiff [74] qui fait suite à Cufflinks [76] pour l'étude de transcrits complets ou encore DEXseq [3], un paquet R qui permet de réaliser l'analyse différentielle de variants d'épissage à l'échelle locale. D'autres méthodes incluent directement l'étape d'analyse différentielle à la suite de leur analyse comme MAJIQ[79] ou Leafcutter[40]. Il est aussi possible de réaliser une analyse différentielle sans génome de référence. À l'échelle locale, KissDE [45] permet de détecter les évènements d'épissage différentiel parmi ceux sortis par KisSplice. KissDE permet aussi d'analyser d'autres évènements d'épissage, détectés et quantifiés en amont par d'autres outils. Pour cela, différents modèles statistiques permettent d'identifier les évènements d'épissage qui sont régulés entre les conditions expérimentales étudiées. Le modèle utilisé dans cette thèse sera décrit plus en détail dans la partie 3. Dans tous les cas, de nombreux tests sont réalisés (un par évènement ou par gène). On sait que plus on réalise de tests statistiques et plus il est probable d'obtenir une p-value faible par hasard. Ainsi, afin de minimiser le nombre de faux positifs obtenus, une étape de correction des p-values pour les tests multiples est indispensable quelle que soit la méthode utilisée. Cette étape est généralement incluse dans le modèle proposé par les méthodes et repose sur la correction de Benjamini-Hochberg [7] ou de Bonferoni.

On s'intéresse aussi à la magnitude de l'effet observé. Pour cela, on calcule des PSI (pour *percent spliced in*). Une valeur de PSI est calculée par condition. Elle correspond à la propor-

tion de transcrits incluant la partie génomique variable (Figure 24). On peut ensuite calculer le ΔPSI, soit la différence des PSI obtenus pour chacune des conditions expérimentales étudiées.

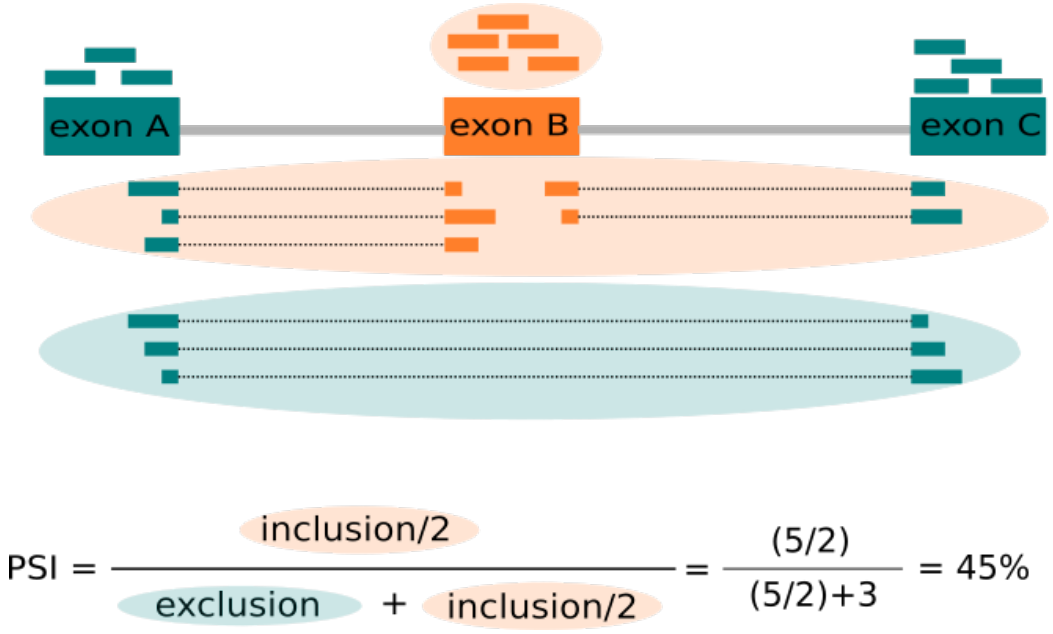


FIGURE 16 – Calcul des valeurs de PSI et ΔPSI. Exemple d'un saut d'exon. L'exon B peut être inclus ou exclu des transcrits. Ici, la partie variable correspond à l'exon B. Les lectures soutenant l'inclusion de l'exon B sont entourées en orange alors que celles soutenant l'exclusion de l'exon B sont entourées en bleue. Le transcrit comprenant l'exon B contient deux jonction alors que celui excluant l'exon B n'en contient qu'une seule. Il est donc nécessaire de diviser par deux le nombre de lectures soutenant les jonctions entre les exons A et B et entre les exons B et C. Les potentielles lectures soutenant les deux jonctions ne sont elles, comptabilisées qu'une seule fois.

4.2 Analyse des données de 3eme génération.

De la même façon que pour les lectures courtes, les analyses réalisées à partir des lectures longues peuvent être basées ou non sur un génome de référence. Les méthodes présentées ici sont plus récentes et l'analyse des lectures longues est un domaine de recherche en plein développement. Même si les technologies de séquençage de troisième génération évoluent rapidement, l'un des enjeux principaux est de prendre en compte le haut taux d'erreur dans les lectures.

Ainsi, il est possible, afin d'obtenir des lectures de meilleure qualité, de les corriger. Certains correcteurs sont hybrides, c'est-à-dire basés sur des lectures courtes dont le taux d'erreur est bien plus faible. Parmi elles, on peut citer HALC [6], LorDEC [65], NaS [47] ou encore PBcR [34]. D'autres outils comme CONSENT [55], Canu [35], ou bien LoRMA [66] permettent de corriger les lectures longues sans l'apport de lectures courtes. Cependant cette étape de correction n'est pas indispensable et il est important de noter qu'elle peut introduire un biais important lorsque l'on s'intéresse à l'épissage : ces méthodes corrigent les lectures, dans la plupart des cas, vers le transcrit majoritaire [41].

Pour l'alignement des longues lectures RNAseq, Minimap2 [39] est l'aligneur le plus utilisé. Cet outil propose un mode permettant d'aligner des lectures épissées sur le génome et ainsi d'ouvrir des gaps suffisamment long, autour des jonctions d'exon. Contrairement aux lectures courtes, qui ne recouvrent généralement qu'une seule jonction d'exon, la plupart des lectures longues en recouvrent plusieurs. Plusieurs gaps peuvent donc être ouverts pour un seul alignement. De plus, Minimap2 permet, suite à une "Feature Request" du consortium ASTER, de prendre en compte une liste de jonctions connues afin de mieux définir les bornes des exons. Il existe d'autre aligneurs adaptés aux lectures longues et aux données RNAseq comme Majic-BLAST [12] ou bien Graphmap2 [51].

En ce qui concerne la quantification des transcrits, des méthodes plus récentes existent.

On peut citer LIQA [26] ou encore Stringtie2 , une méthode hybride qui, en combinant l'utilisation des lectures courtes et des lectures longues, permet de reconstruire des transcrits complets et de les quantifier [36]
UNAGI [1] et FLAIR , sont deux pipelines qui permettent l'identification et la quantification des transcrits complets depuis l'alignement des lectures longues sur un génome de référence.

Concernant les approches sans génome de référence, puisque les lectures sont longues et peuvent parfois couvrir des transcrits complets, les méthodes d'assemblage ne sont pas adaptées. On cherche alors à regrouper entre elles les lectures qui pourraient provenir d'un même transcrit (clustering). Cela permet d'identifier les différents transcrits présents dans les données mais aussi de les quantifier en comptant le nombre de lectures obtenues dans chaque groupe. Ces méthodes, comme RATTLE [14] ou CARNAC-LR [49], prennent en compte le taux élevé d'erreur de séquençage dans les données.

Deuxième partie

Objectifs de la thèse

Ma thèse s'inscrit dans le travail du consortium ASTER (https ://bioinfo.cristal.univ-lille.fr/aster/). Celui-ci a pour but d'évaluer et de développer des méthodes d'analyse de données issues des technologies de séquençage de 3ème génération. Les collaborations avec l'équipe RD du Genoscope et plus particulièrement avec Jean-Marc Aury et Corine Da Silva ont permis de produire des données que j'ai analysé dans une première partie de ma thèse. Je me suis interrogée sur les apports de ces technologies pour l'étude des transcriptomes eucaryotes et plus particulièrement sur la fiabilité des quantifications obtenues avec les données Nanopore.

Ensuite je me suis intéressée à la modélisation des évènements d'épissage complexes. KisSplice [64] est un assembleur local de transcriptome développé dans l'équipe BAOBAB dans laquelle j'ai réalisé ma thèse. Cette méthode permet de détecter et de quantifier des évènements d'épissage alternatif sans génome de référence. À mon arrivée dans l'équipe, KisSplice ne permettait d'étudier que des évènements d'épissage simple, c'est-à-dire de comparer des transcrits deux à deux. Cependant, on sait que les variations d'épissage peuvent être plus complexes et parfois plus de trois transcrits peuvent êtres produits localement. Il n'existe pour l'instant aucune méthode publiée permettant l'étude des évènements complexes sans génome de référence. J'ai donc cherché à adapter le modèle de KisSplice aux évènements complexes pour les cas où le modèle pairwise n'était pas suffisant. J'ai à coeur que mes travaux s'inscrivent dans une problématique biologique et il me paraissait donc important de travailler sur la façon de modéliser les évènements d'épissage complexes avant de traduire le problème en problème mathématique. Je voulais proposer un modèle réaliste, basé sur mes connaissances des mécanismes moléculaires liés à la réaction d'épissage. Le but est d'aider les biologistes intéressés par des gènes cibles à faciliter la compréhension des variations d'épissage observées et à mettre en évidence des arguments pour discuter des questions encore ouvertes concernant l'épissage et sa régulation. J'ai donc cherché des exemples de gènes soumis à des variations d'épissage complexes afin de tester mes hypothèses puis la pertinence des résultats obtenus une fois la méthode implémentée.

Troisième partie

Apport des données de troisième génération pour l'étude des transcriptomes

1 | Contexte

Dans une première partie de ma thèse, j'ai analysé des données RNAseq de 3ème génération. Les différents jeux de données ont été produits au Genoscope, à ÉVRY, dans l'équipe "RD bio-informatique et séquençage". Les interactions au sein du consortium ASTER et en particulier avec Jean-Marc Aury et Corinne Da Silva nous ont permis d'avoir accès à des jeux de données récents. Ceux-ci ont été produits selon des protocoles expérimentaux et des technologies tout juste commercialisés au moment du séquençage. Cela nous a notamment permis d'évaluer la qualité d'un des premiers jeux de données issus des protocoles de RNA direct proposé par Nanopore qui permet de séquencer les molécules d'ARN directement sans étape de rétro-transcription en ADN complémentaire. Nous avons également pu comparer ce jeu de données à des données de troisième génération plus classique (protocole ADNc de Nanopore). Un troisième protocole a été testé : le protocole Teloprime. Celui-ci permet d'enrichir les données en transcrits complets (polyadénylés et cappés). Enfin, nous avons pu comparer ces jeux de données avec des données de 2ème génération que nous avions plus l'habitude d'utiliser. En effet, des échantillons commerciaux de cerveau et de foie de souris ont été séquencés avec les technologies Illumina et Nanopore (RNAdirect, cDNA et teloprime). Dans le cas du cerveau, plusieurs réplicats ont été générés afin de pouvoir tester la reproductibilité de nos résultats.

Nous avons cherché à savoir comment les données issues des technologies de 3ème génération pouvaient nous aider à étudier les transcriptomes eucaryotes. Plus particulièrement après avoir étudié les caractéristiques globales (profondeur, taux d'erreur, taille des lectures) des différents jeux de données nous nous sommes demandés si les quantifications des gènes

et des transcrits obtenues avec ces jeux de données étaient fiables. Nous les avons comparées avec les quantifications obtenues avec Illumina (au niveau des gènes ou des transcrits) que nous considérons généralement comme précises. Afin d'avoir une mesure objective de la fiabilité de ces quantifications nous avons aussi utilisé des spike-ins. Ce sont des transcrits artificiels dont les quantifications sont connues. Nous disposions du Mix E2 des spikes-ins Lexogen. Ce mix est composé de 67 transcrits et, contrairement aux mix E0 et E1, les quantifications des différents transcrits varient de façon importante. Nous savions donc combien de transcrits étaient présents dans les librairies avant les étapes de séquençage et nous avons donc pu regarder si nous retrouvions ces quantifications dans les données obtenues. Afin d'évaluer l'effet du protocole de préparation des librairies nous avons aussi séquencé des lectures courtes (Illumina), en utilisant le protocole cDNA de Nanopore (en ajoutant une étape de fragmentation). Cela nous a permis de mettre en évidence que le protocole RNAdirect de Nanopore ainsi que les données Illlumina permettent d'obtenir les quantifications les plus précises pour quantifier les transcrits de chacun des gènes.

Je me suis aussi rendu compte, en étudiant les quantifications obtenues pour les gènes après alignement sur le génome de référence, que pour certaines populations de gènes j'obtenais des quantifications bien plus élevées avec les longues lectures qu'avec les lectures Illumina. En regardant de plus près des exemples de ces gènes puis en généralisant le processus en attribuant à chacun le "gene biotype" qui lui est associé dans les annotations, je me suis rendue compte qu'il s'agissait, en très grande majorité, de pseudogènes rétrotranscrits. J'ai d'abord cru que ceux-ci étaient plus exprimés que ce que nous pensions (ils sont connus pour être très peu exprimés [28]). Cette hypothèse a généré un enthousiasme important, notamment lors de mon premier comité de suivi de thèse. Je me suis finalement rendue compte, après plusieurs semaines et grâce à des discussions constructives avec les membres de mon comité mais aussi avec d'autres membres du laboratoire que ces quantifications élevées étaient dues à des erreurs d'alignements. En effet les pseudogènes rétrotranscrits ne possèdent pas d'intron et il est plus facile d'aligner les lectures longues en un seul bloc plutôt que d'ouvrir des gaps importants. De plus, la présence de queue polyA dans les lectures ainsi que dans les

pseudogènes rétrotranscrits (mais pas dans les gènes) amplifie ce biais. Des exemples de pseudogènes, sur lesquels des lectures longues étaient alignées avec des mismatchs correspondant exactement aux divergences connues avec le gène parent m'ont permis de m'en convaincre (Figure 6 du papier ci-dessous). Pour résoudre ce problème, nous avons proposé d'aligner les longues lectures sur le transcriptome de référence plutôt que sur le génome.

Même si les erreurs et les résultats négatifs ne sont habituellement pas présentés dans nos disciplines il me paraissait important de décrire les étapes par lesquelles je suis passée pour obtenir ce résultat. Cela pourrait permettre à d'autres de ne pas reproduire cette erreur.

Un papier récent [78] affirme, après analyse de données de lectures longues, que certains pseudogènes rétrotranscrits sont finalement exprimés. Nous avons contacté les auteurs, afin de mieux comprendre leur protocole bioinformatique et leur réponse soulève de nouvelles questions complémentaires. À ce jour, la question n'est pas tranchée et je ré-analyse une partie de leurs données (ré-alignements sur le transcriptome notamment) afin de comprendre si leurs résultats sont basés sur la même erreur que celle que j'avais faite dans un premier temps ou non.

2 | Publication dans Scientific Reports (2019)

Un article à ce sujet à été publié dans Scientific Reports en Octobre 2019.

SCIENTIFIC REPORTS
nature research

OPEN

Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules

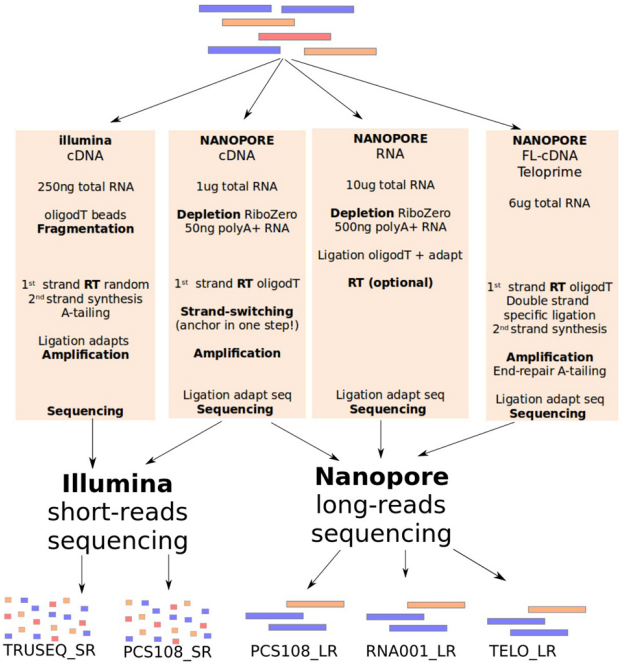
Camille Sessegolo[1,4], Corinne Cruaud[2], Corinne Da Silva[2], Audric Cologne[1,4], Marion Dubarry[2], Thomas Derrien[3], Vincent Lacroix[1,4] & Jean-Marc Aury[2*]

Our vision of DNA transcription and splicing has changed dramatically with the introduction of short-read sequencing. These high-throughput sequencing technologies promised to unravel the complexity of any transcriptome. Generally gene expression levels are well-captured using these technologies, but there are still remaining caveats due to the limited read length and the fact that RNA molecules had to be reverse transcribed before sequencing. Oxford Nanopore Technologies has recently launched a portable sequencer which offers the possibility of sequencing long reads and most importantly RNA molecules. Here we generated a full mouse transcriptome from brain and liver using the Oxford Nanopore device. As a comparison, we sequenced RNA (RNA-Seq) and cDNA (cDNA-Seq) molecules using both long and short reads technologies and tested the TeloPrime preparation kit, dedicated to the enrichment of full-length transcripts. Using spike-in data, we confirmed that expression levels are efficiently captured by cDNA-Seq using short reads. More importantly, Oxford Nanopore RNA-Seq tends to be more efficient, while cDNA-Seq appears to be more biased. We further show that the cDNA library preparation of the Nanopore protocol induces read truncation for transcripts containing internal runs of T's. This bias is marked for runs of at least 15T's, but is already detectable for runs of at least 9T's and therefore concerns more than 20% of expressed transcripts in mouse brain and liver. Finally, we outline that bioinformatics challenges remain ahead for quantifying at the transcript level, especially when reads are not full-length. Accurate quantification of repeat-associated genes such as processed pseudogenes also remains difficult, and we show that current mapping protocols which map reads to the genome largely over-estimate their expression, at the expense of their parent gene.

To date our knowledge of DNA transcription is brought by the sequencing of RNA molecules which have been first reverse transcribed (RT). This RT step is prone to skew the transcriptional landscape of a given cell and erase base modifications. The sequencing of these RT-libraries, that we suggest to call cDNA-Seq, has became popular with the introduction of the short-read sequencing technologies[1,2]. Recently, the Oxford Nanopore Technologies (ONT) company commercially released a portable sequencer which is able to sequence very long DNA fragments[3] and enable now the sequencing of complex genomes[4–6]. Moreover this device (namely MinION) is also able to sequence native RNA molecules[7] representing the first opportunity to generate genuine RNA-Seq data.

Furthermore, even if short-read technologies offer a deep sequencing and were helpful to understand the transcriptome complexity and to improve the detection of rare transcripts, they still present some limitations. Indeed, read length is a key point to address complex regions of a studied transcriptome. Depending on the evolutionary history of a given genome, recent paralogous genes can lead to ambiguous alignment when using short reads. In a same way, processed pseudogenes generated by the retrotranscription of RNAs back into the genomic DNA are challenging to quantify using short reads. In addition to sequencing technologies and bioinformatics methods,

[1]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622, Villeurbanne, France. [2]Genoscope, Institut de biologie François-Jacob, Commissariat a l'Energie Atomique (CEA), Université Paris-Saclay, F-91057, Evry, France. [3]Univ Rennes, CNRS, IGDR (Institut de génétique et développement de Rennes) - UMR 6290, F-35000, Rennes, France. [4]EPI ERABLE - Inria Grenoble, Rhône-Alpes, France. *email: jmaury@genoscope.cns.fr



Brain sample 1	C1R1	X	X	X	X	X
Brain sample 2	C2R1	X		X	X	
	C2R2	X		X	X	
Liver Sample 1	C1R1	X	X	X	X	X

Figure 1. Experimental design. Five protocols have been used on each tissue. Two were based on short-reads with the TruSeq protocol (TRUSEQ_SR) and the ONT library preparation (PCS108_SR) and the three others were based on long-reads with the ONT cDNA-Seq protocol (PCS108_LR), the ONT RNA-Seq protocol (RNA001_LR) and the Teloprime protocol (TELO_LR). (RT: Reverse Transcription). For the brain, two biological replicates, C1 and C2, have been generated and two technical replicates, R1 and R2, have been generated for the second biological replicate. For the first biological replicate all the five protocols were used whereas the TRUSEQ_SR, the PCS108_LR and the RNA001_LR were used for the second biological replicates.

preparation protocols have a significant impact on the final result as they can incorporate specific biases[8,9]. The generation of data rely on a high number of molecular and computational steps which evolve at a fast pace. These changes in the protocol generally modified the appearance of the data. As an example, data produced with protocols based on oligo-dT or random primers in the RT step show differences in how they cover transcripts[8].

Results

Experimental design. Here we produce a complete transcriptome dataset, containing both cDNA-Seq and RNA-Seq, using the Illumina and Nanopore technologies. RNAs were sampled from brain and liver tissues of mice and were mixed with Lexogen's Spike-In RNA Variants (SIRVs) as a control for quantification of RNAs. We follow the protocols recommended by the manufacturers to generate the three following datasets on each tissue: Illumina cDNA-Seq, Nanopore cDNA-Seq and Nanopore RNA-Seq. The first was sequenced using the Illumina platform (TruSeq_SR) and the last two using the MinION device (PCS108_LR and RNA001_LR). From the brain tissue, we generated biological (two brain RNA samples, C1 and C2) and technical replicates (R1 and R2) for the three datasets (Fig. 1). Additionally, the second was also sequenced using the Illumina platform (PCS108_SR). This enables us to clarify which differences are due to the preparation protocol and which are due to the sequencing platform in itself. Moreover, we generated a Lexogen's TeloPrime library on both tissues (TELO_LR), this preparation kit is an all-in-one protocol for generating full-length cDNA from total RNA (Fig. 1 and Tables 1 and 2).

Spliced alignment and error rate. The error rate of ONT reads is still around 10% and complicates the precise detection of splice sites. Mouse splice sites are often canonical, as observed when aligning reference annotation (coding genes from Ensembl 94) using BLAT, 98.5% of introns were GT-AG. Here minimap2 was able to detect only 80.7% of GT-AG introns when ONT RNA-Seq reads were used as input. Interestingly, the proportion of cannonical splice sites is lower when using ONT cDNA-Seq (67.7%). In fact ONT RNA-Seq reads are strand-specific which is of high value for the alignment and splice site detection. When using high quality sequences (coding genes from Ensembl 94) instead of reads, minimap2 retrieved 96.4% of GT-AG introns. These results show that the detection of splice sites using long but noisy reads is challenging and that dedicated aligners still need some improvements.

RNA sample	PCS108_LR			RNA001_LR			TELO_LR
	C1	C2		C1	C2		C1
Replicate	R1	R1	R2	R1	R1	R2	R1
	Brain						
Number of reads	1,267,830	5,834,882	3,003,844	571,098	364,041	210,654	1,691,454
Cumulative size (Gb)	1.30	7.03	3.28	0.43	0.38	0.20	1.31
Average Size (bp)	1,028.94	1,204.54	1,091.85	758.05	1,032.10	957.17	775.77
N50 (bp)	1,283	1,749	1,591	1,357	1,492	1,417	896
Number reads >1 Kb	522,422	2,869,633	1,339,489	154,735	141,970	73,232	389,468
Accession number	ERX2695238	ERX3387950	ERX3387952	ERX2695236	ERX3387949	ERX3387951	ERX2850744
	Liver						
Number of reads	3,043,572	—	—	418,102	—	—	2,668,975
Cumulative size (Gb)	3.30	—	—	0.34	—	—	2.64
Average Size (bp)	1,083.85	—	—	823.30	—	—	989.85
N50 (bp)	1,264	—	—	1,153	—	—	1,116
Number of reads >1 Kb	1,218,569	—	—	117,047	—	—	884,008
Accession number	ERX2695243	—	—	ERX2695240	—	—	ERX2850745

Table 1. Standard metrics of the generated ONT datasets for both tissues: brain and liver.

RNA sample	TruSeq_SR			PCS108_SR
	C1	C2		C1
Replicate	R1	R1	R2	R1
	Brain			
Number of reads	53,128,934	41,562,993	45,719,216	153,610,181
Cumulative size (Gb)	15.42	12.36	13.60	45.88
Read Size (bp)	151	151	151	151
Accession number	ERX2695239	ERX3387947	ERX3387948	ERX2695237
	Liver			
Number of reads	49,270,153	—	—	178,019,939
Cumulative size (Gb)	14.16	—	—	53.43
Read Size (bp)	151	—	—	151
Accession number	ERX2695241	—	—	ERX2695242

Table 2. Standard metrics of the Illumina datasets for both tissues: brain and liver.

General comparison of sequencing technologies. RNA-Seq is a powerful method that provides a quantitative view of the transcriptome with the number of sequenced fragments being a key point to thoroughly capture the expression of genes. The Illumina technology is able to generate billions of short tags, and unsurprisingly allows to access a largest number of genes/transcripts. However with the same number of reads, both Illumina and Nanopore technology are able to uncover the same number of transcripts (Fig. 2). Long reads sequencing offers the possibility to capture full-length RNAs. When looking at single isoform genes, we found that in average reads cover between 61 and 74% of the messenger RNAs (Table 3). But even though this horizontal coverage is quite high, the proportion of reads that covered more than 80% of the transcript remains low (near 55% except for cDNA-Seq of the brain sample). RNA degradation can obviously explain a proportion of these fragmented reads, and it has been shown more recently that a software artifact may truncate reads (around 20%) during the base-calling process[10].

Improving the proportion of full-length reads. The proportion of full-length transcripts can be improved by using a dedicated library preparation protocol. Here we tested the TeloPrime amplification kit, commercialized by the Lexogen company. This protocol is selective for full-length RNA molecules that are both capped and polyadenylated. Using this protocol we were able to slightly improve the proportion of full-length reads (which covered at least 80% of a given transcript, Table 3). However, in return, we captured a lower number of genes, lowering the interest of such a protocol in most applications (Fig. 2). The TeloPrime amplification kit is no longer available in this form. A new version is now available and it will be interesting to see if it brings improvements.

Sequencing biases of transcripts containing internal runs of poly(T). Since cDNA synthesis is initiated with an anchored poly-dT primer (poly-TVN), a relevant question is whether transcripts containing internal runs of poly(A) or poly(T) are correctly sequenced. We computed the relative coverage for each transcript upstream and downstream internal runs of poly(A) or poly(T) (see Methods), and we find that using cDNA-Seq,

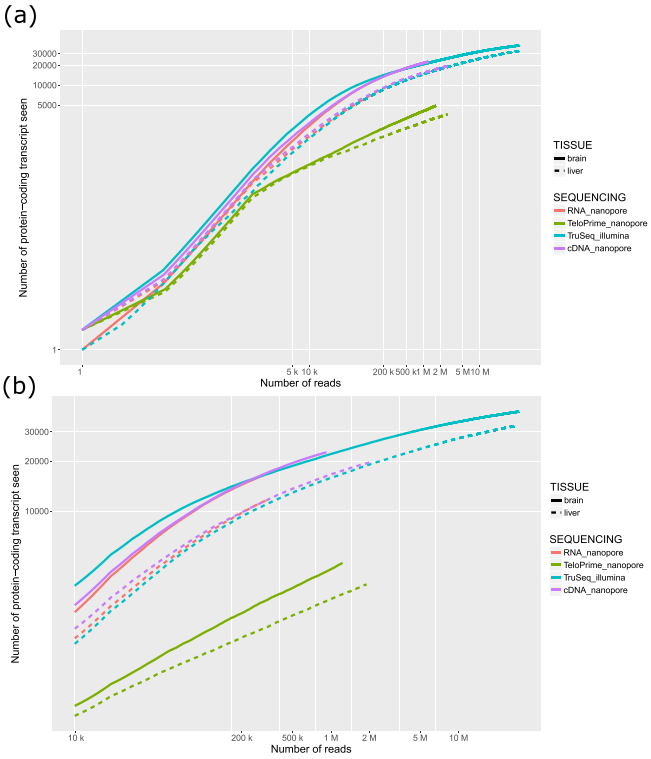


Figure 2. Saturation curve. Number of protein-coding transcripts seen by each technology at various sequencing depth. Solid and dashed lines correspond respectively to brain and liver samples. (**a**) Both axes starting from 1 (**b**) zoom of the graph above, masking the regime from 1 to 10k reads which is less interesting when comparing technologies as it does not correspond to any real-world experiment.

	PCS108_LR			RNA001_LR			TELO_LR
RNA sample	**C1**	**C2**		**C1**	**C2**		**C1**
Replicate	**R1**	**R1**	**R2**	**R1**	**R1**	**R2**	**R1**
	Brain						
# of mapped reads	99,192	389,007	203,622	45,485	48,128	29,719	200,884
Avg coverage	61.18%	62.18%	62.10%	70.97%	74.25%	76.52%	76.37%
Median coverage	64.62%	62.11%	61.59%	83.84%	89.34%	92.36%	84.65%
Full-length reads (>80%)	40.25%	41.08%	40.73%	53.31%	57.19%	60.61%	60.18%
	Liver						
# of mapped reads	344,362	—	—	48,032	—	—	381,005
Avg coverage	73.86%	—	—	73.48%	—	—	79.09%
Median coverage	83.24%	—	—	84.57%	—	—	88.23%
Full-length reads (>80%)	55.63%	—	—	54.77%	—	—	60.02%

Table 3. Long reads coverage of single-isoform genes.

cDNA molecules stemming from such transcripts are often truncated. This bias is detectable for runs of poly(A) (Fig. 3b) but is much stronger for runs of poly(T) (Fig. 3a). While the first situation could be caused by internal poly(T) priming during first strand cDNA synthesis and therefore result in 3′-truncation of the cDNA, the second situation could occur during 2nd strand cDNA synthesis and result in 5′-truncation of the cDNA (as sketched in Fig. 3c). As an example, the *Set* gene contains an internal run of 20 T's and ONT cDNA-Seq reads are systematically interrupted at this location (Fig. 3c, tracks 2 and 3), while this is not the case for Illumina Truseq (Fig. 3d, track 4) and ONT RNA-Seq (Fig. 3d, track 1). We find that the magnitude of the bias is associated to the length of the internal run of poly(T) (Supplementary Fig. 1). The bias is very pronounced for transcripts containing at least 15 T's, but it is already detectable for transcripts containing at least 9 T's. This bias has remained unreported so far, but it is also present in other published Nanopore dataset[11] (Supplementary Fig. 2). It however concerns a large fraction of expressed transcripts. Indeed, transcripts containing at least 9 T's correspond to 27% of transcripts expressed with at least one read in mouse brain (resp. 20% in mouse liver). In human GM12878 cell line, this proportion is 16%. Importantly, the bias not only affects read length, but also transcript quantification. Indeed,

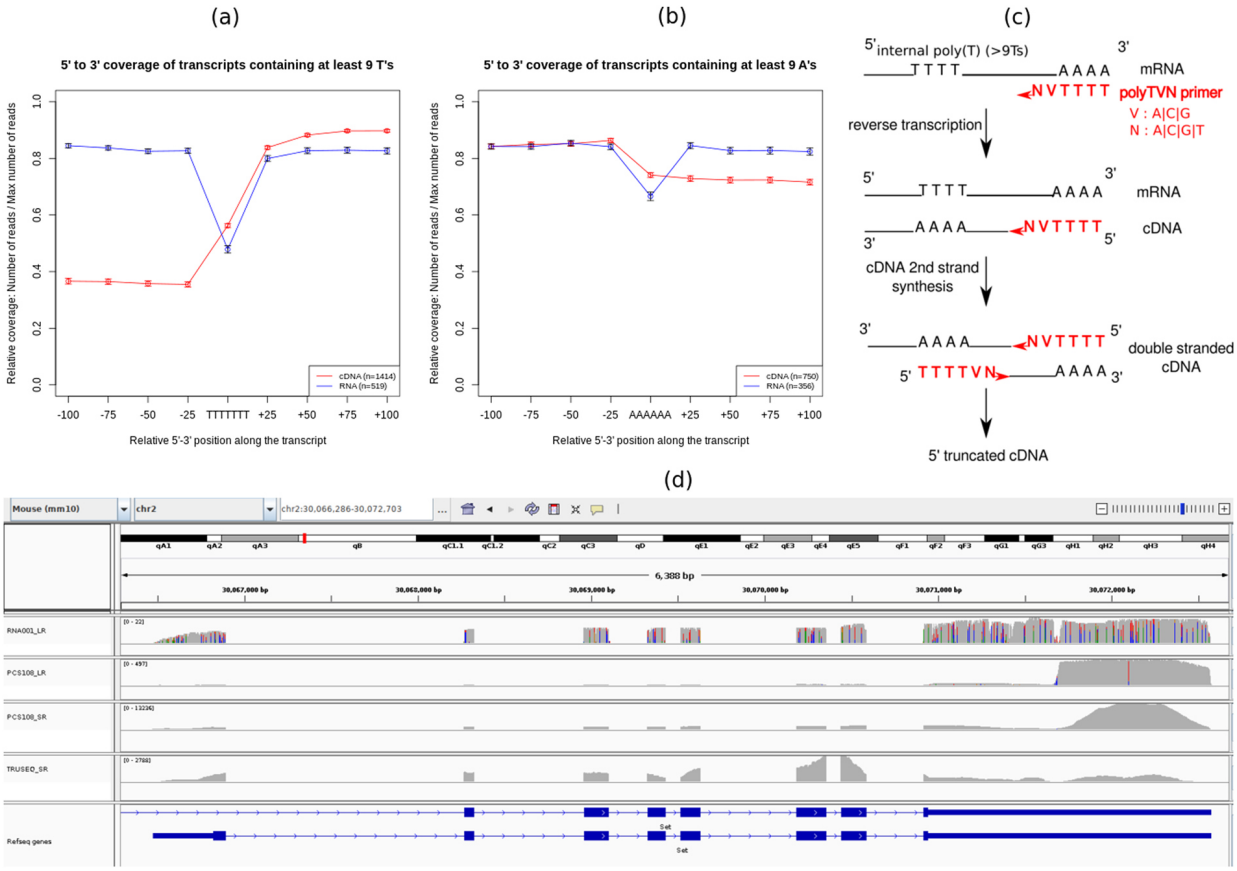


Figure 3. Truncated reads. (**a**) Relative coverage of transcripts for the ONT cDNA-Seq dataset and the ONT RNA-Seq dataset for transcripts covered by at least 10 reads around a poly(T). With the ONT cDNA-Seq dataset, transcripts containing internal runs of at least 9 T's are less covered in 5′. The coverage deficit observed in the ONT RNA-seq dataset is due to indel sequencing errors associated to homopolymers. (**b**) Relative coverage of transcripts for the ONT cDNA-Seq dataset and the ONT RNA-Seq dataset for transcripts covered by at least 10 reads around a poly(A). Using the ONT cDNA-Seq dataset, transcripts containing stretches of at least 9 A's are less covered in 3′. Again, the coverage deficit observed in the ONT RNA-seq dataset is due to indel sequencing errors associated to homopolymers. (**c**) Mechanism explaining why internal runs of T's are causing 5′ truncated reads. The PolyTVN primer binds to the internal run of poly(A) of the cDNA so that the second cDNA strand is 5′ truncated. (**d**) Example of a gene named *Set* visualized with IGV. Truncated reads are in tracks 2 (ONT cDNA-Seq) and 3 (Illumina, Nanopore protocol). Non-truncated reads are in tracks 1 (ONT RNA-Seq) and 4 (Illumina Truseq). The region where the truncation occurs is a poly(T).

cDNA-Seq reads from these transcripts are not only shorter, they are also more numerous. As an example, the *set* gene is covered by 497 truncated cDNA reads and 22 full-length RNAseq reads (Fig. 3d). More generally, in mouse brain, 35% of cDNA-Seq reads map to transcripts with at least 9 T's, compared to 14% of RNA-Seq reads. This suggests that the abundance of these transcripts is over-estimated when using cDNA-Seq, at the expense of the other transcripts.

Evaluation of the accuracy of the gene expression quantification using spike-in data. In order to assess which protocol was best to quantify gene expression, we analyzed the 67 spike-ins contained in the brain datasets. Since we exactly know which transcripts are present in the sample, the quantification is rather straightforward. We aligned reads to the reference transcriptome, used RSEM for short reads, and counted the number of primary alignments for long reads (see Methods). The best quantifications were obtained for the ONT RNA-Seq (Spearman $\rho = 0.86$, Pearson $r = 0.85$) and Illumina TruSeq ($\rho = 0.81$, $r = 0.82$) protocols (Fig. 4). In contrast, cDNA-Seq (sequenced using Illumina or ONT) produced more imprecise quantifications ($\rho = 0.54$, $r = 0.57$ and $\rho = 0.6$, $r = 0.50$). Importantly, we obtained very similar results on all our three replicates (Supplementary Figs 3 and 4), with ONT RNA-Seq consistently exhibiting the higher correlation with true quantifications. The use of salmon either for short or long reads, as was done in[12] did not change our results. We then wanted to test if the number of ONT RNA-Seq reads was indeed a better predictor of the true transcript quantification, than the number of cDNA-Seq reads or the TPM measure derived from Illumina. Using 30 fold cross-validations, we found that the mean square error was $ms_{cDNA} \in [1.55; 1.57]$ for ONT cDNA-Seq, $ms_{illu} \in [1.35; 1.41]$ for Illumina and $ms_{rna} \in [0.76; 0.78]$ for ONT RNA-Seq. When inspecting the errors made for each SIRV, we noticed that SIRV311 was particularly poorly predicted by all methods (possibly because it is only 191nt long which makes it the

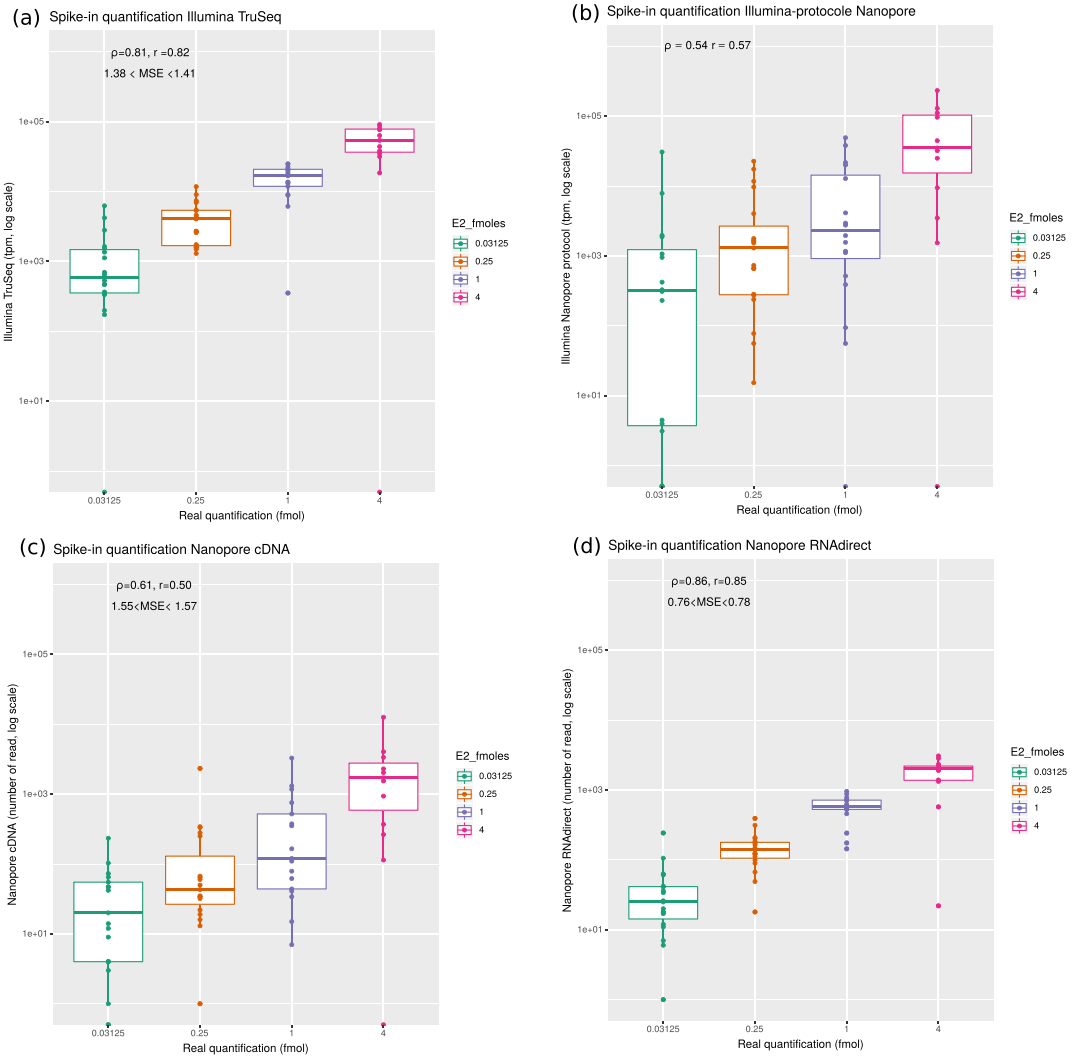


Figure 4. Evaluation of quantification using the SIRV E2 spike-in mix. Reads were mapped against the SIRV transcriptome and quantifications were computed at the transcript level. The observed quantifications are correlated with the known theoretical quantifications of the spike in. in. Mean square error (MSE) were computed using a cross validation approach (see methods). (**a**) Correlation obtained for Illumina cDNA-Seq (Spearman's $\rho = 0.80$, n = 67 transcripts). (**b**) Correlation obtained for Illumina with the ONT cDNA-Seq protocol (Spearman's $\rho = 0.53$, n = 67 transcripts). (**c**) Correlation obtained for ONT cDNA-Seq (Spearman's $\rho = 0.65$, n = 67 transcripts). (**d**) Correlation obtained for ONT RNA-Seq (Spearman's $\rho = 0.86$, n = 67 transcripts).

shortest SIRV), and in particular by Illumina TruSeq. When removing it from the dataset, we obtained $ms_{illu} \in [0.623; 0.634]$, $ms_{rna} \in [0.483; 0.498]$, $ms_{cDNA} \in [1.44; 1.47]$ which highlights that ONT RNA-Seq yields significantly better quantifications than Illumina TruSeq and ONT cDNA-Seq. Although the magnitude of the difference with Illumina TruSeq is small, we found it to be reproducible. We could further show that, for each technology, the errors made for each SIRV were reproducible across replicates (Supplementary Fig. 5) meaning that a transcript whose expression is over-estimated with one technology is consistently over-estimated with the same technology.

In order to assess the quality of the quantification in a more realistic context where we do not know which transcripts are present in the sample, we also mapped the reads to a modified set of transcripts corresponding either to an over-annotation or an under-annotation (as provided by Lexogen). In both cases, the correlations were overall poorer than before, but the order was maintained, with ONT RNA-Seq and then Illumina cDNA-Seq being the more reliable protocols (Supplementary Figs 6 and 7).

Quantification of the expression level of mouse transcripts. Given that with the spike in, the best quantification were obtained with ONT RNA-Seq, we compared the quantifications obtained with this protocol with the ones obtained with the ONT and Illumina cDNA-Seq protocols. Figure 5 summarizes the correlations in terms of transcript quantification of our datasets. Comparing the Illumina cDNA-Seq and the ONT RNA-Seq protocols we obtain a spearman coefficient of correlation $\rho = 0.51$. The correlation is lower in the liver sample (Supplementary Fig. 8) probably because of a lower number of RNA-seq reads and a shorter read length.

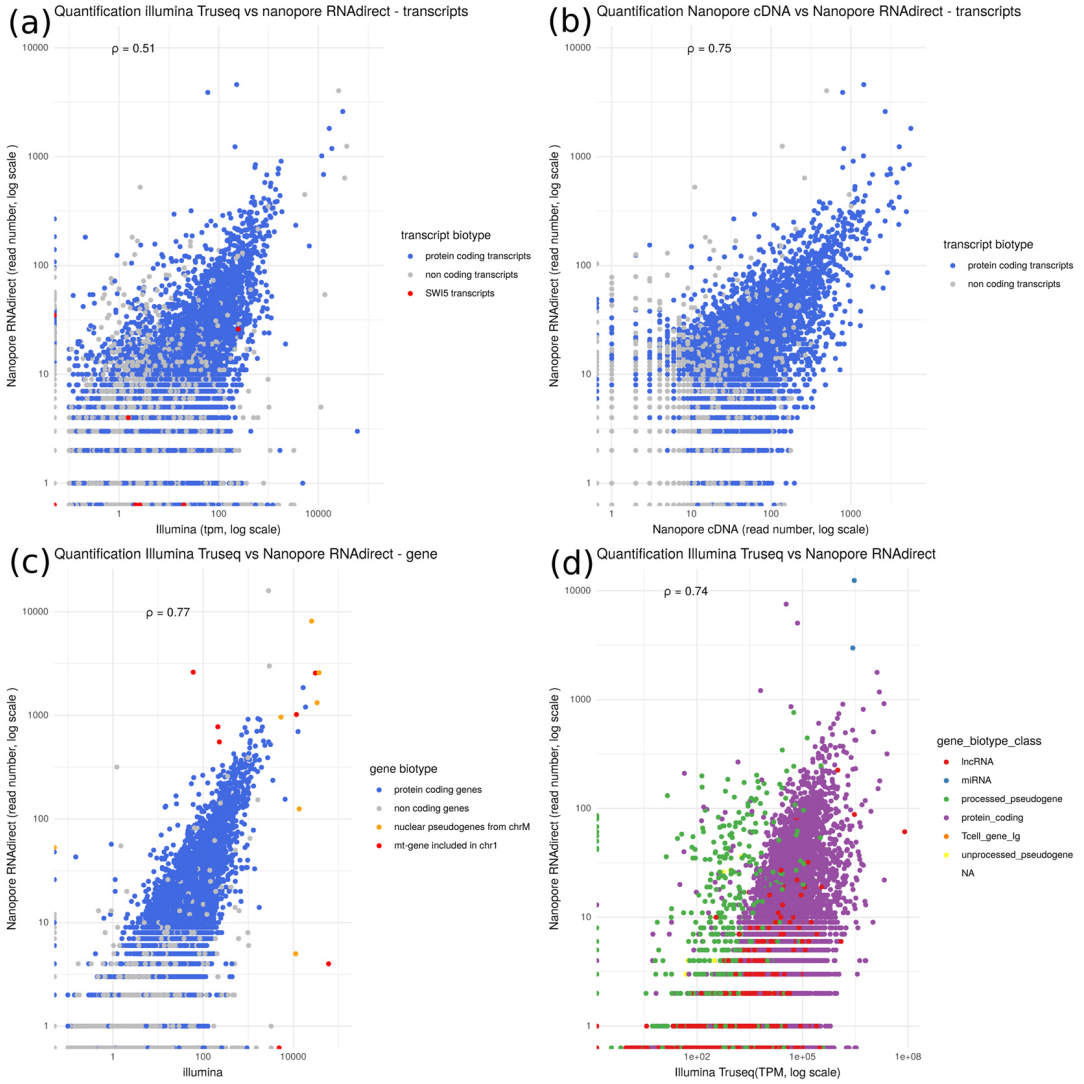


Figure 5. Comparison of quantifications. Transcripts or genes annotated as protein-coding are in blue. Spearman's ρ has been computed for all transcripts or genes. (**a**) Comparison of ONT RNA-Seq and Illumina cDNA-Seq quantifications at the transcript level (Spearman's $\rho = 0.51$, n = 140,325). Red points correspond to the transcripts of the *Swi5* gene. (**b**) Comparison of ONT RNA-Seq and ONT cDNA-Seq quantifications at the transcript level (Spearman's $\rho = 0.75$, n = 140,325). (**c**) Comparison of ONT RNA-Seq and Illumina cDNA-Seq quantifications at the gene level (transcript quantification were summed for each gene, Spearman's $\rho = 0.77$, n = 54,532). Red points correspond to pseudogenes located on chromosome 1 within the NUMT, i.e. segment of the mitochondrial genome which has been copied and integrated in the nuclear genome and orange points correspond to the original mitochondrial genes. (**d**) Reads were mapped against the mouse reference genome and quantifications computed at gene level. We compared the ONT RNA-Seq and the Illumina cDNA-Seq protocols (Spearman's $\rho = 0.74$, n = 54,532). Green points correspond to processed pseudogenes, red points to long non coding RNAs.

Comparing the ONT RNA-Seq and cDNA-Seq quantification, we obtain a higher correlation ($\rho = 0.75$), suggesting that read length strongly influences transcript quantification. Indeed, in the comparison between Illumina cDNA-Seq and ONT RNA-Seq dataset, the lack of correlation comes from one main cause. Discriminating transcripts of a same gene that share common sequences with short reads is difficult. Longer reads are clearly helpful, however they do not always enable to discriminate transcripts. Indeed, in the case where a read only covers the 3′ end of a transcript, and not the full length, it may be ambiguously assigned to several transcripts.

For example, for the *Swi5* gene, although several rare (lowly expressed) transcripts are seen only with Illumina, the other ones are harder to quantify (red dots in Fig. 5a). RSEM uses the unique part of each transcript to proportionally allocate the reads that mapped equally on the common part of the transcript. In the case where a transcript has no read which uniquely maps to it, its expression cannot be computed and is set to 0. This is the case for the transcript ENSMUST00000050410 (*Swi5*-201, Supplementary Fig. 9) of *Swi5*, whose expression is underestimated (0 TPM). Conversely, some transcripts are underestimated by ONT RNA-Seq. This is the case of *Swi5*-204, whose unique region is located at the 5′ end of the gene, and is therefore poorly covered by long reads.

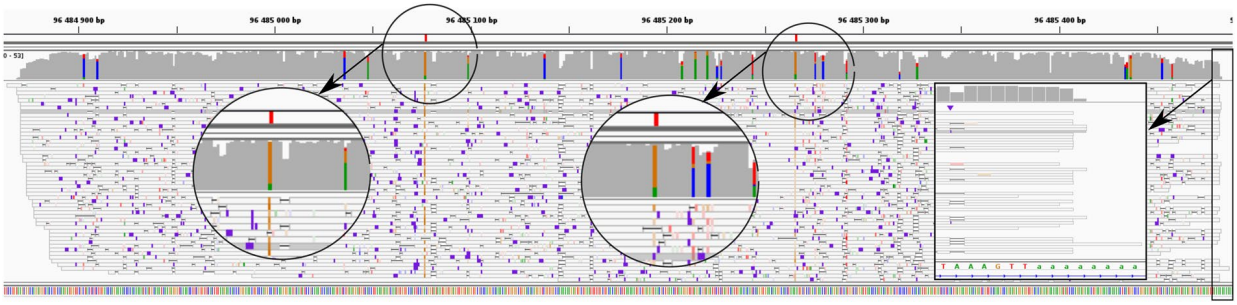


Figure 6. Example of a processed pseudogene whose expression is overestimated: *Rpl17-ps8* (retro Rpl17) Alignment visualization with IGV of *Rpl17-ps8*. The positions of divergence between Rpl17-ps8 and Rpl17 are shown in red in the first track. Second track is ONT RNA-Seq coverage, third track is ONT RNA-Seq reads. Colored positions in the coverage track correspond to mismatches. Most reads contain mismatches at the exact position of the divergences with the parent gene. They are therefore incorrectly mapped, partly because they overlap the polyA tail which is integrated in the genome downstream the pseudogene.

To avoid the difficult step of correctly assigning a read to a transcript, we summed the quantification of all transcripts for each gene. Figure 5c shows the quantification at gene level. As reported in other papers[7,13,14] the correlation at gene level is quite good ($\rho = 0.78$). However inter-genes repeats remains a cause of mis-quantified genes. For example, a large part of the mitochondrial chromosome had been recently integrated in the mouse chromosome 1[15]. As a consequence, 7 genes are present in 2 copies in the genome, one copy annotated as functional on the mitochondrial chromosome and another one, annotated as pseudogene on chromosome 1 (shown in red in Fig. 5c). Since this integration is recent, the copies did not diverge yet. They are therefore difficult to quantify due to multimapping, even when using long reads, since the repeat is larger than the full transcript.

Quantification of processed pseudogenes. These are particular cases of processed pseudogenes which come from the retrotranscription and reintegration in the genome of one of the transcript of their parent gene[16]. After their integration, they have no intron and, without any selective pressure, they diverge from their parent gene proportionally with their age. Some of them are expressed[17] and are annotated as transcribed processed pseudogenes although the vast majority of pseudogenes are not expressed[16]. Correctly assigning the reads to the parent gene and not the pseudogene is not trivial.

Figure 5d shows that mapping long reads to the genome with Minimap 2 (-ax splice) (as used in[11]) results in the mis-quantification of processed pseudogenes (green points in Fig. 5d). The expression of most of them is over-estimated by the ONT RNA-Seq protocol (this is also the case of ONT cDNA-Seq). It can be explained by two main reasons.

First, it can come from the fact that if a mapper has to choose between two genomic locations, one with gaps (introns of the parent gene), and one with no gaps (the processed pseudogene), it will tend to favour the gapless mapping, as gapless alignment are easier to find. We note that the scoring system of minimap2 consists in selecting the max-scoring sub-segment, excluding introns, and therefore not explicitly favouring the gapless mapping. However, this requires that splice sites are correctly identified in the first place, a task which remains difficult with noisy long reads.

A second reason explaining the overestimation of processed pseudogenes is related to polyA tails. Processed pseudogenes originate from transcripts which contained a polyA tail, which was then integrated in the genome, downstream the pseudogene. Many of the ONT reads originating from the parent gene also contain this polyA tail, favoring the alignment at the processed pseudogene genomic location. The alignment will be longer thanks to the polyA tail. An example is shown in Fig. 6. The processed pseudogene *Rpl17-ps8* differs from its parent gene by two bases (A to G at the position chrX:96,485,078 and A to G at the position chrX:96,485,267). These divergences are marked in red in the figure. At these two positions we observe that reads differ from the reference genome: they have a G instead of an A. This means that these reads come from the parent gene and we mistakenly aligned them onto the pseudogene because it is intronless and contains a polyA tail.

Quantification of genes overlapping transposable elements. Another example of repeat-associated gene biotype is given by the long non-coding RNAs (lncRNAs) which are highly enriched in transposable elements (TEs). These TEs are sometimes considered as the functional domains of lncRNAs[18], and it has been estimated that 66% of mouse lncRNA overlap at least one TE[19]. Our experimental design allows us to assess the impact of read lengths on non-coding (versus protein-coding) gene quantifications for different levels of TE coverage. As expected, the higher the TE content of a gene, the larger the difference in quantification between long and short-read sequencing technologies (Supplementary Fig. 10a). Although this tendency is observed for both protein-coding and lncRNAs biotypes, lncRNAs are more impacted given that they are more prone to be enriched in TEs. One interesting example is given by the known imprinted lncRNA *KCNQ1OT1* (ENSMUSG00000101609) which is specifically expressed from the paternal allele in opposite direction to the *KCNQ1* protein-coding gene[20] (Supplementary Fig. 10b). About 41% of the *KCNQ1OT1* transcript sequence is composed of TE elements and its quantification using Illumina TruSeq versus ONT cDNA-Seq protocols highlights contrasting values (TPM = 0.36 and ONT cDNA-Seq = 118).

Discussion

In this work, we generated a dataset which we think should be of general interest for the community. This dataset consists of RNA and cDNA sequencing of the same samples using both Illumina and ONT technologies. Importantly, we also sequenced Lexogen E2 spike-in data, together with our mouse samples, which enabled us to assess which technology yielded the most accurate quantification.

Although lexogen spike-in have been used to evaluate the quantification obtained with ONT cDNA-Seq[21] or ONT RNA-Seq[7] protocol separately, we are the first to compare the quantification obtained with ONT cDNA-Seq, RNA-Seq and Illumina cDNA-Seq.

Using the spike-in data, we find that the ONT RNA-Seq protocol is the most accurate, slightly better than the widely used Illumina TruSeq protocol. In contrast, the cDNA-Seq data was more biased and yielded a poorer quantification.

We further found that transcripts with internal runs of poly(T) tend to be truncated and over-sampled when using the ONT cDNA-Seq protocol. Sequencing the same library preparation with the Illumina technology enabled us to confirm that the truncation issue was related to the sample preparation and not to the sequencing. We further show that this bias is not restricted to our dataset, and can be found in a human ONT dataset[11]. Truncation biases associated to internal runs of poly(A) had been reported earlier and motivated the usage of anchored poly-dT primers (poly-TVN)[22]. On the other hand, biases associated to internal runs of poly(T) had remained undetected, although they may affect more than 20% of expressed transcripts in mouse. This bias could also affect other long-reads cDNA-Seq data. Although biases had been searched for in previous work[23], it may have remained undetected because the authors were then focusing on internal runs of at least 20 A's.

We then used our data to quantify mouse genes and found that ONT RNA-Seq quantification correlated well with Illumina cDNA-Seq quantification (Fig. 5c) but when trying to quantify at the transcript level, the correlation was overall poorer (Fig. 5a). A temptation could be to think that ONT RNA-Seq yields better transcript-level quantification as reads are longer and are, unlike short reads, unambiguously assigned to a single transcript. In practice, 70% of ONT RNA-Seq reads are assigned to a single transcript, while the remaining 30% are ambiguously mapped. This was particularly the case for transcripts which differed at their 5′end, like in Swi5. Quantifying transcripts and not genes is still challenging, and requires the development of dedicated bioinformatics methods. When trying to use salmon[24] on long reads, as in[12], we did not obtain better results than when simply counting primary alignments. There should however be room for improvement in this direction, and our spike-in dataset could be a good training set for future methods.

In this work, we chose to align reads to a reference transcriptome. Indeed, when trying to map reads to the reference genome, we observed a systematic over-estimation of the quantification of processed pseudogenes, at the expense of their parent gene. We further show that this biased quantification is due to alignment issues: 1- poly(A) tails of pseudogenes are integrated in the genome and'attract' reads from the parent gene and 2- accurate identification of splice sites when mapping long RNA-Seq reads is challenging, which disfavors the parent gene.

We therefore strongly recommend to map reads on the reference transcriptome and not on the genome, as reference transcripts do not contain introns, nor poly(A) tails. However, a clear limitation of aligning reads to a reference annotation, instead of a reference genome, is that we cannot discover novel transcripts. As a consequence, reads stemming from these novel transcripts will be unmapped, or incorrectly assigned to alternative transcripts (as in APOE gene, Supplementary Fig. 11). Improving alignment tools to correctly handle processed pseudogenes seems essential to identify and quantify transcripts, especially in the case of non-model species where no exhaustive annotation is available.

More generally, the quantification of repeat-containing genes is difficult. Long reads are particularly useful for quantifying these genes, like long non coding RNAs, which are enriched in transposable elements.

There is currently a lot of interest for the potential of ONT RNA-Seq to identify and quantify genes and transcripts, as can be seen by the currently low but expanding number of datasets available with this technology. Here we proposed the first dataset on mouse with several interesting and unique features, as Lexogen E2 spike-ins, Illumina sequencing of ONT library preparation or Lexogen TeloPrime protocol. We think that ONT sequencing is promising for studying RNA, especially if the number of reads and full-length reads continues to increase. Improvements in the technology and library preparation protocol to obtain more reads and more full-length reads are also expected to be very helpful in obtaining precise quantification of all genes and transcripts. The recent launch of the PromethION device will allow a deep sequencing of transcriptomes which should enable to overcome the limitations of the MinION device.

Methods

Biological material. We used total RNA extracted from mouse brain (Cat #636601, lot number 1403636A and 1605262A) and liver (Cat #636603, lot number 1305118A) from Clontech (Mountain View, CA, USA).

Libraries preparation. *Illumina cDNA library*. RNA-Seq library preparations were carried out from a mix of 250 ng total RNA and 0.25 ng Spike-in RNA Variant Control Mix E2 (Lexogen, Vienna, Austria) using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA), which allows mRNA strand orientation. Ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems, Wilmington, MA, USA), and libraries profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

Illumina on Nanopore cDNA library. 250 ng of cDNA prepared using the "DNA-PCR Sequencing" protocol (see "Nanopore cDNA library" below) were sonicated to a 100- to 1000-bp size using the E220 Covaris instrument (Covaris, Woburn, MA, USA). Fragments were end-repaired, then 3′-adenylated, and NEXTflex PCR free

barcodes adapters (Bioo Scientific, Austin, TX, USA) were added using NEBNext Sample Reagent Module (New England Biolabs, Ipswich, MA, USA). Ligation products were amplified using Illumina adapter-specific primers and KAPA HiFi Library Amplification Kit (KapaBiosystems, Wilmington, MA, USA) and then purified with AMPure XP beads (Beckmann Coulter, Brea, CA, USA). Ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems), and libraries profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies).

Nanopore cDNA library. Total RNA was first depleted using the Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina). RNA was then purified and concentrated on a RNA Clean Concentrator™-5 column (Zymo Research, Irvine, CA, USA). cDNA libraries were performed from a mix of 50 ng RNA and 0.5 ng Spike-in RNA Variant Control Mix E2 (Lexogen) according to the Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) protocol "DNA-PCR Sequencing" with a 14 cycles PCR (8 minutes for elongation time). ONT adapters were ligated to 650 ng of cDNA.

Nanopore RNA library. RNA libraries were performed from a mix of 500 ng RNA and 5 ng Spike-in RNA Variant Control Mix E2 (Lexogen) according to the ONT protocol "Direct RNA sequencing". We performed the optional reverse transcription step to improve throughput, but cDNA strand was not sequenced.

Nanopore TeloPrime library. Three cDNA libraries were performed from 2 μg total RNA for each RNA sample according to the TeloPrime Full-Length cDNA Amplification protocol (Lexogen). A total of 5 PCR were carried out with 30 to 40 cycles for the brain sample and 30 cycles for the liver sample. Amplifications were then pooled and quantified. Nanopore libraries were performed from respectively 560 ng and 1000 ng of cDNA using the SQK-LSK108 kit according to the Oxford Nanopore protocol.

Sequencing and reads processing. *Illumina datasets.* Illumina cDNA libraries, prepared with the TruSeq (TruSeq_SR) and Nanopore (PCS108_SR) protocols, were sequenced using 151 bp paired end reads chemistry on a HiSeq4000 Illumina sequencer (Table 1). After the Illumina sequencing, an in-house quality control process was applied to the reads that passed the Illumina quality filters. The first step discards low-quality nucleotides ($Q < 20$) from both ends of the reads. Next, Illumina sequencing adapters and primer sequences were removed from the reads. Then, reads shorter than 30 nucleotides after trimming were discarded. The last step identifies and discards read pairs that mapped to the phage phiX genome, using SOAP[25] and the phiX reference sequence (GenBank: NC_001422.1). These trimming and removal steps were achieved using in-house-designed software as described in[26].

Nanopore datasets. Nanopore libraries were sequenced using a MinION Mk1b with R9.4.1 (PCS108_LR and RNA001_LR) or R9.5 flowcells (TELO_LR). The data were generated using MinKNOW 1.11.5 and basecalled with Albacore 2.1.10 (PCS108_LR C1R1, RNA001_LR C1R1 and TELO_LR C1R1) or MinKNOW 3.1.19 and basecalled with Guppy 2.3.5 (PCS108_LR C2R1 and C2R2, RNA001_LR C2R1 and C2R2, see Table 1).

Reads alignment and transcripts quantification. Long reads were mapped to the spike-in transcripts using Minimap2 (version 2.14)[27] (-ax map-ont). Supplementary alignments, secondary alignments and reads aligned on less than 80% of their length were filtered out. We used the number of aligned reads as a proxy of the expression of a given transcript. Short reads were mapped to the spike-in transcripts using bowtie[28] and quantified using RSEM[29]. The quantification obtained is given in TPM (transcript per million).

We then assessed the mouse transcripts expression and mapped the long reads against the mouse transcripts (Ensembl 94) using Minimap2 (with the following options -ax map-ont and -uf for direct RNA reads). Long reads from cDNA (PCS108_LR) and TeloPrime (TELO_LR) were trimmed using porechop and default parameters before alignment against the mouse transcripts. Long reads from RNA (RNA001_LR) were not trimmed, as the ONT basecaller could not detect DNA adapters. Supplementary alignements, secondary alignments and reads aligned on less than 80% of their length were filtered out. Expression was directly approximated by the number of reads which mapped on a given transcript. Long reads were also mapped on the reference genome using Minimap2 (-ax splice). Supplementary alignments, secondary alignments and reads aligned on less than 80% of their length were filtered out. Short reads were mapped to the reference genome (release Grcm38.p6) using STAR with the gtf option (annotation Ensembl 94). In order to quantify each transcript, short reads were also mapped on the referencence transcriptome using bowtie and quantification were obtained with RSEM.

Evaluating the ability of each technology to predict the true SIRV quantification using cross-validation. We build 3 models: M_1: $\log(SIRV) = \mu_1 + \beta_1 * \log(readCountcDNA) + error$; M_2: $\log(SIRV) = \mu_2 + \beta_2 * \log(TPM) + error$; M_3: $\log(SIRV) = \mu_3 + \beta_3 * \log(readCountRNA) + error$. As these models are not nested, they cannot be compared against each other with likelihood ratio tests. We therefore use cross-validation, using 4/5 of our 67 SIRV to estimate the parameters of each model, and the remaining 1/5 to estimate the quality of the prediction. We repeat this process 30 times, choosing randomly a different partition to train and test the model, and we obtain confidence intervals on the prediction error for each model.

Truncated reads analysis. For each transcript annotated in Ensembl94 containing an internal run of at least 9Ts, we computed the number of reads covering the following positions: 25, 50, 75 and 100nt upstream and downstream the internal run of poly(T). For each transcript *t*, the most covered position was retrieved, and the number of reads covering this position was noted max_t. The coverage of each position was then divided by max_t,

so as to obtain a normalised coverage. Then for each position, we computed the mean of the relative coverage at this position across all transcripts verifying $max_t > 10$. This is the value plotted in Fig. 3a. The error bars represent the standard error around the mean. The same analysis was done for the human ONT dataset, using gencode27 annotations (Supplementary Fig. 2).

Saturation curve. For short reads, we kept only the best alignment as reported by RSEM and the primary alignment of each long read. Only protein coding transcripts (transcript_biotype = protein_coding) were taken into account.

Quantification of TE-containing genes. Given that lncRNAs are lowly expressed, for this specific analysis, we restricted to cDNA-Seq and did not apply our 80% query coverage filter. Using annotated TEs from the RepeatMasker database[30], we classified lncRNAs and mRNAs based on their TE coverage in four categories (with the "0%" class corresponding to genes without any exonic-overlapping TE and conversely, the class of ">66–100%" for genes highly enriched in exonic TE). For each expressed gene, we further computed the ratio between Nanopore cDNA versus Illumina TruSeq gene quantifications with respect to their TE categories.

Data availability

The Illumina and MinION data are available in the European Nucleotide under the following accession number PRJEB27590. The entire dataset (fastq and bam files) is available from the following website: http://www.genoscope.cns.fr/ont_mouse_rna/.

Received: 20 March 2019; Accepted: 28 September 2019;
Published online: 17 October 2019

References

1. Lipson, D. *et al*. Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnology* **27**, 652–658, issn: 1087-0156 (July 2009).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57–63, issn: 1471-0064 (Jan. 2009).
3. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* **34**, 518–524, issn: 1087-0156 (May 2016).
4. Belser, C. *et al*. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* **4**, 879–887, issn: 2055-0278 (Nov. 2018).
5. Jain, M. *et al*. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338–345, issn: 1087-0156 (Jan. 2018).
6. Schmidt, M. H.-W. *et al*. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant cell* **29**, 2336–2348, issn: 1532-298X (Oct. 2017).
7. Garalde, D. R. *et al*. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods* **15**, 201–206, issn: 1548-7091 (Jan. 2018).
8. Alberti, A. *et al*. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15**, 912, issn: 1471-2164 (Oct. 2014).
9. Van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* **322**, 12–20, issn: 00144827 (Mar. 2014).
10. Brooks, A. *(Nanopore RNA Consortium) - Native RNA sequencing of human polyadenylated transcripts*, https://nanoporetech.com/resource-centre/native-rna-sequencing-human-polyadenylated-transcripts [Accessed 25 Fev 2019] (2018).
11. Workman, R. E. *et al*. Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*, https://doi.org/10.1101/459529, eprint: https://www.biorxiv.org/content/early/2018/11/09/459529.full.pdf, https://www.biorxiv.org/content/early/2018/11/09/459529 (2018).
12. Soneson, C. *et al*. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *bioRxiv*, https://doi.org/10.1101/574525, eprint: https://www.biorxiv.org/content/early/2019/03/11/574525.full.pdf, https://www.biorxiv.org/content/early/2019/03/11/574525 (2019).
13. Byrne, A. *et al*. ARTICLE Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications* **8**, https://doi.org/10.1038/ncomms16027, https://www.nature.com/articles/ncomms16027.pdf (2017).
14. Seki, M. *et al*. Evaluation and application of RNA-Seq by MinION. *DNA Research*, dsy038 (2018).
15. Leister, D. & Richly, E. NUMTs in Sequenced Eukaryotic Genomes. *Molecular Biology and Evolution* **21**, 1081–1084, issn: 0737-4038 (June 2004).
16. Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics* **10**, 19–31, issn: 1471-0064 (Jan. 2009).
17. Carelli, F. N. *et al*. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome research* **26**, 301–14, issn: 1549-5469 (Mar. 2016).
18. Johnson, R. & Guigo, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20**, 959–976, issn: 1355-8382 (July 2014).
19. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13**, R107, issn: 1465-6906 (2012).
20. Mancini-DiNardo, D., Steele, S. J. S., Levorse, J. M., Ingram, R. S. & Tilghman, S. M. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes & Development* **20**, 1268–1282, issn: 0890-9369 (May 2006).
21. Weirather, J. L. *et al*. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100, issn: 2046-1402 (Feb. 2017).
22. Nam, D. K. *et al*. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6152–6, issn: 0027-8424 (Apr. 2002).
23. Kuo, R. I. *et al*. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323, issn: 1471-2164 (Apr. 2017).
24. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (Apr. 2017).

25. Li, R. *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967, issn: 1367-4803 (Aug. 2009).
26. Alberti, A. *et al*. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data* **4**, 170093, issn: 2052-4463 (Aug. 2017).
27. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
28. Langmead, B. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics* Chapter 11, Unit 11.7, issn: 1934-340X (Dec. 2010).
29. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, issn: 1471-2105 (Dec. 2011).
30. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* Chapter 4, Unit 4.10, issn: 1934-340X (Mar. 2009).

Acknowledgements

The authors are grateful to Oxford Nanopore Technologies Ltd for providing early access to the MinION device through the MAP, and we thank the staff of Oxford Nanopore Technology Ltd for technical help, particularly Botond Sipos, Daniel Turner, Michelle Hiscutt and James Platt for insightful discussions about poly-T truncation phenomenon. We thank Philippe Veber and Arnaud Mary for their advice on statistical and bioinformatics analyses. This work was supported by French National research agency (ANR project ANR-16-CE23-0001 'ASTER'), the INRIA, the Genoscope, the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and France Génomique (ANR-10-INBS-09-08).

Author contributions

C.C. optimized and performed the sequencing. C.S., C.D.S., A.C., M.D., T.D., V.L. and J.M.A. performed the bioinformatic analyses. C.S., V.L. and J.M.A. wrote the article. V.L. and J.M.A. supervised the study.

Competing interests

The authors declare that they have no financial competing interests. C.C. and J.M.A. are part of the MinION Access Programme (MAP) and J.M.A. received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-51470-9.

Correspondence and requests for materials should be addressed to J.-M.A.

Reprints and permissions information is available at www.nature.com/reprints.

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

3 | Matériel supplémentaire à la publication précédente

Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules

Camille Sessegolo[1,4], Corinne Cruaud[2], Corinne Da Silva[2], Audric Cologne[1,4], Marion Dubarry[2], Thomas Derrien[3], Vincent Lacroix[1,4], Jean-Marc Aury[2,*]

[1] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558 F-69622 Villeurbanne, France.
[2] Genoscope, Institut de biologie François-Jacob, Commissariat a l'Energie Atomique (CEA), Université Paris-Saclay, F-91057 Evry, France
[3] Univ Rennes, CNRS, IGDR (Institut de génétique et développement de Rennes) - UMR 6290, F-35000, Rennes,France
[4] EPI ERABLE - Inria Grenoble, Rhône-Alpes, France.
[*] Correspondence to jmaury@genoscope.cns.fr

Supplementary Figures

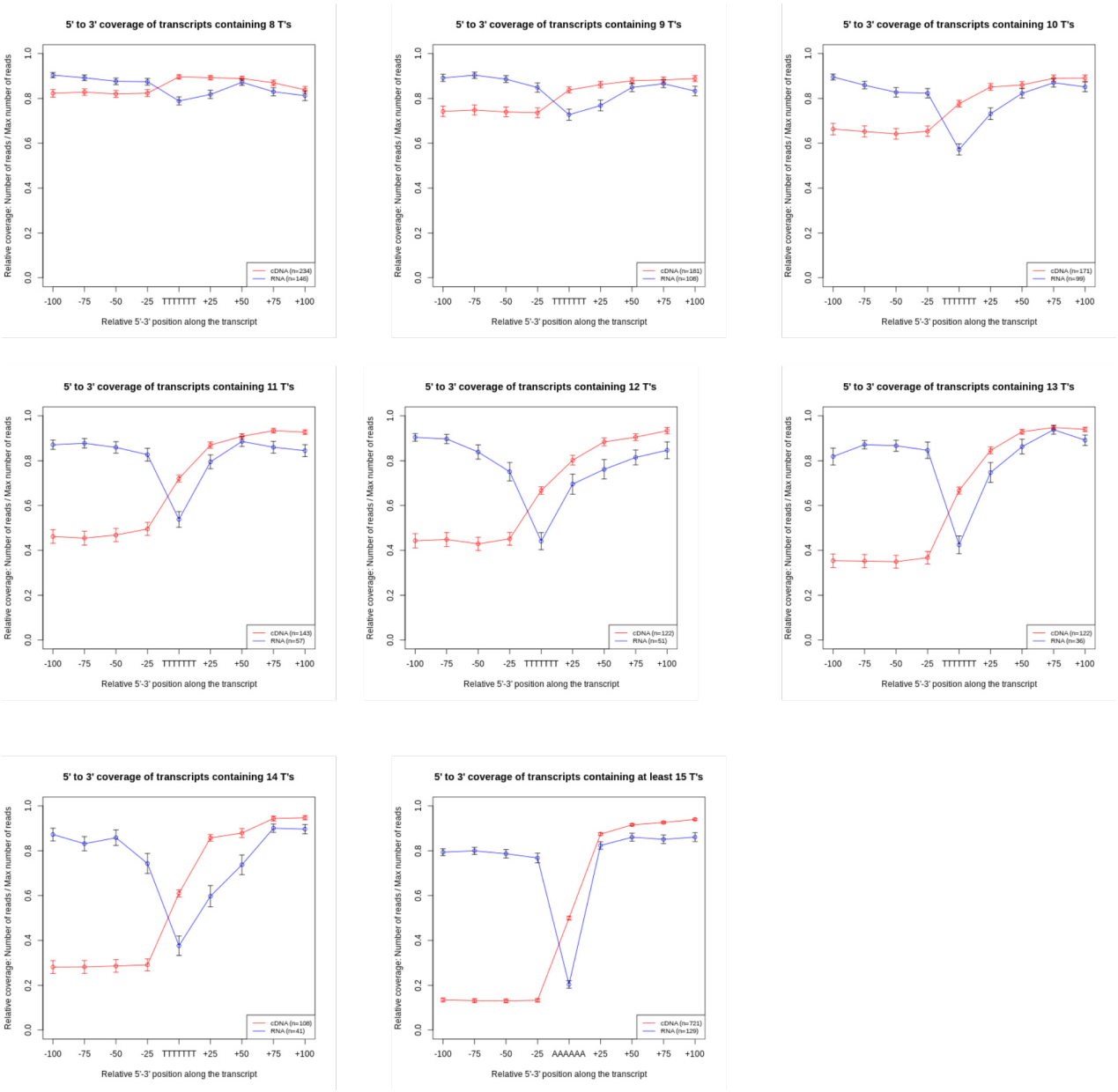


Figure 1: **Poly(T) induce 5' truncated reads** Relative coverage of transcripts for our ONT cDNA-Seq dataset and our ONT RNA-Seq dataset for transcripts covered by at least 10 reads around a poly(T). Several size of poly(T) have been tested and we found that the effect is visible using the cDNA-Seq dataset from poly(T) longer than 9 T's : transcripts containing stretches of at least 9 T's are less covered in 5' than other transcripts. In all cases, the local coverage deficit observed in the ONT RNA-seq dataset is due to sequencing error causing by the homopolymers.

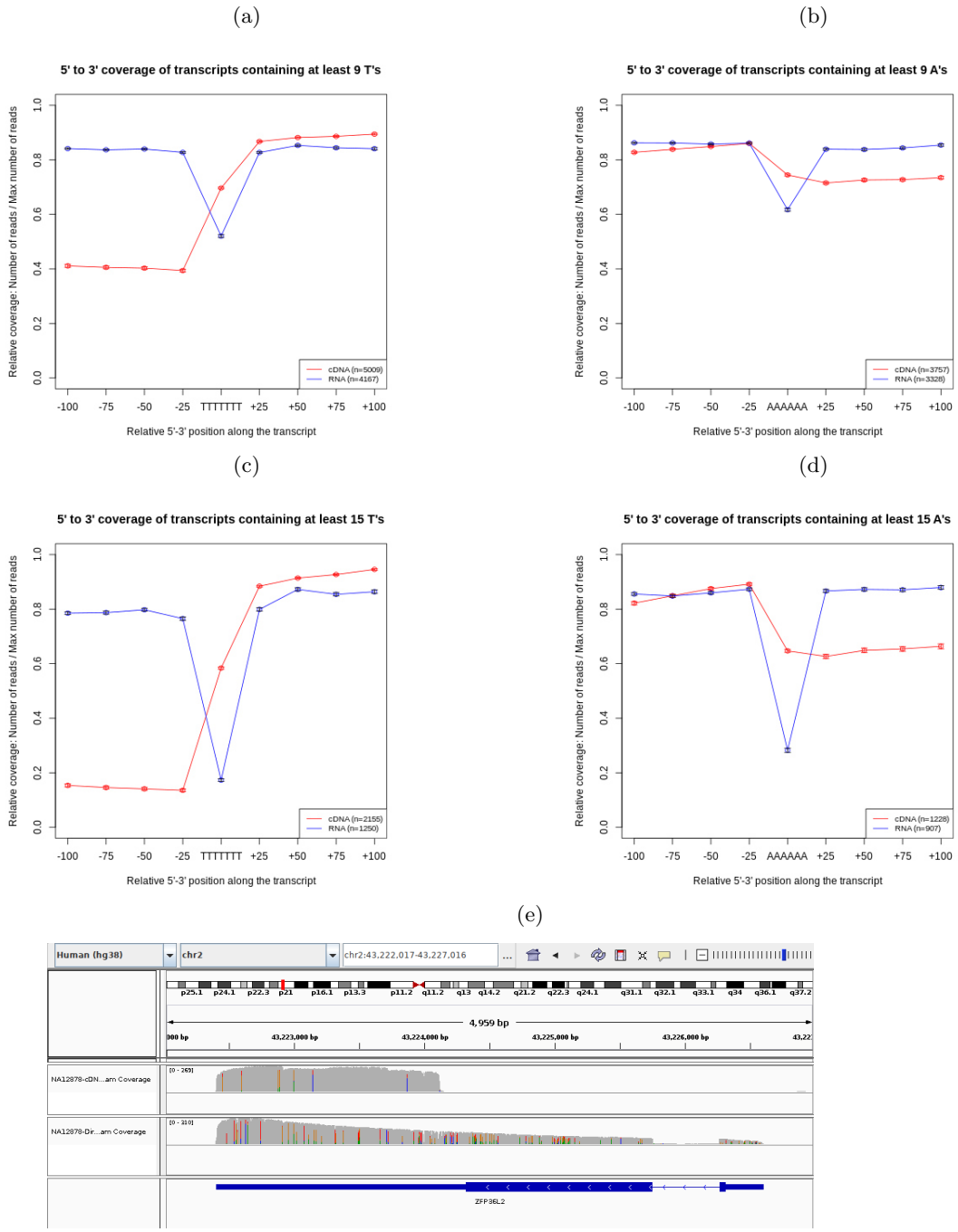


Figure 2: **Truncated Reads** Relative coverage of transcripts for the ONT cDNA-Seq dataset and the ONT RNA-Seq dataset of Workman et al. for transcripts covered by at least 10 reads around an internal run of poly(T) (panels a and c) or poly(A) (panels b and d). Using the ONT CDNA-Seq dataset, transcripts containing internal runs of poly(T) are less covered in 5' than other transcripts, wheras transcripts containing internal runs of poly(A) are less covered in 3'. It indicates that these transcripts are covered by a high proportion of truncated reads. The coverage deficit observed in the ONT RNA-seq dataset is due to sequencing errors caused by the homopolymers. The effect is more marked when considering internal runs of at least 15 T's (panel c) or 15 A's (panel d). (e) Example obtained using the Workman et al. dataset. The *ZFP36L2* gene contains an internal run of 11 T's. Reads from the ONT cDNA-Seq are truncated (first track) whereas ONT RNA-Seq reads are not (second track).

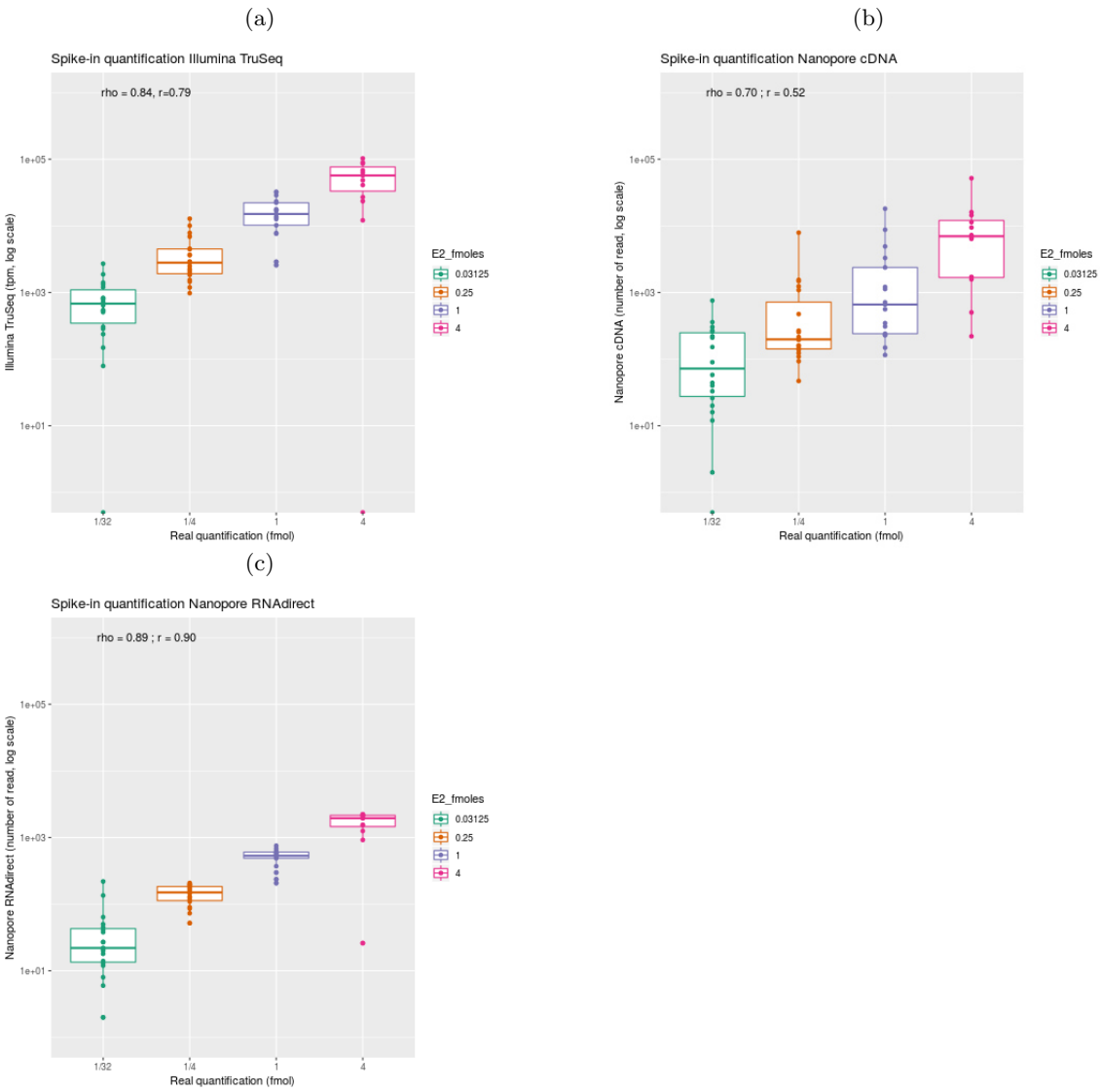


Figure 3: **Spike in quantification for C2R1 datasets** Reads were mapped against the SIRV transcriptome and quantifications computed at transcript level. The observed quantification are correlated with the known theoretical quantification of the spike in. (a) Correlation obtained for Illumina with the TruSeq protocol (Spearman's rho =0.84). (b) Corelation obtained for Nanopore with the cDNA protocol (Spearman's rho = 0.70). (c) Corelation obtained for Nanopore with the RNA direct protocol (Spearman's rho = 0.89).

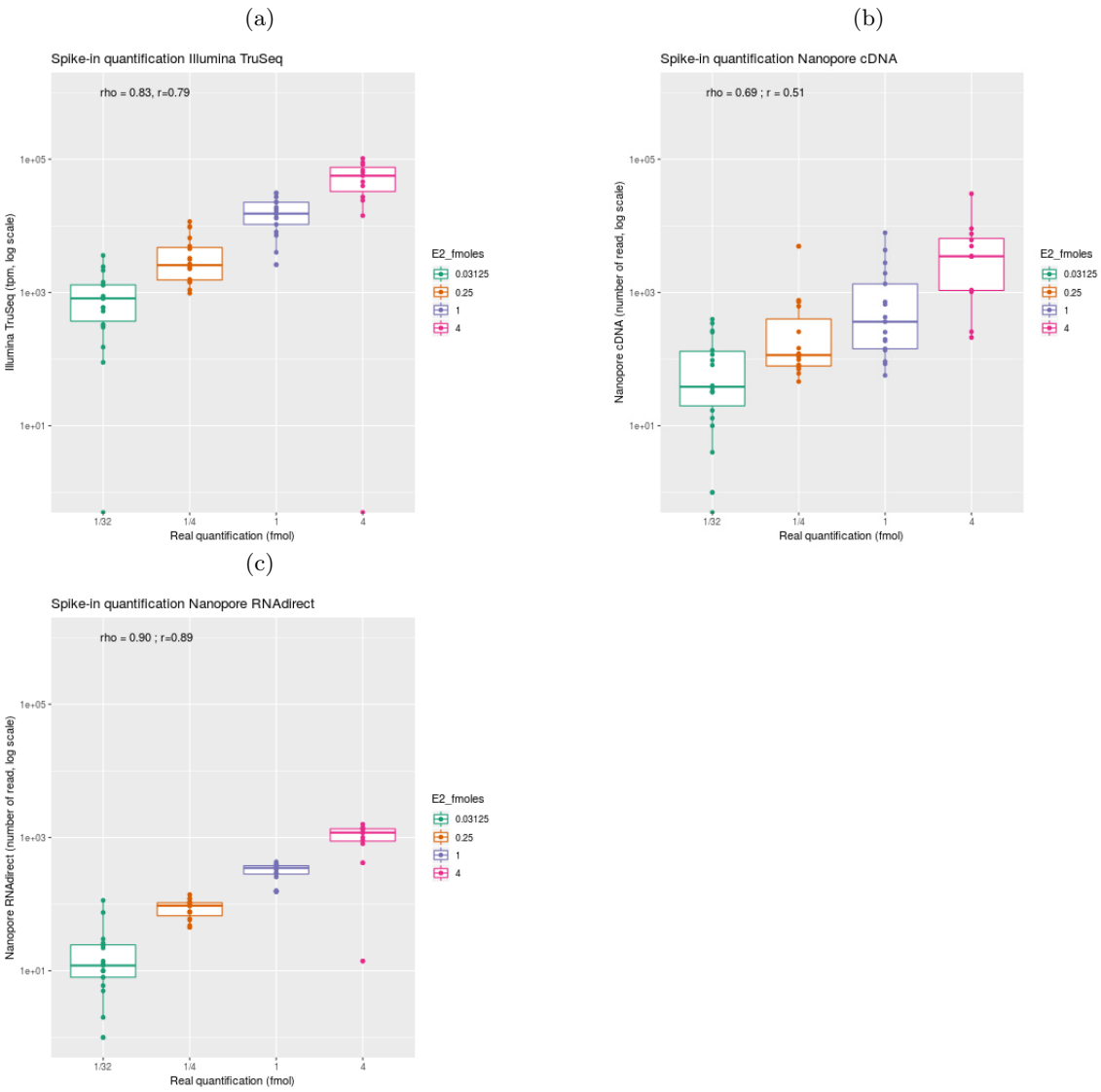


Figure 4: **Spike in quantification for C2R2 datasets** Reads were mapped against the SIRV transcriptome and quantifications computed at transcript level. The observed quantification are correlated with the known theoretical quantification of the spike in. (a) Correlation obtained for Illumina with the TruSeq protocol (Spearman's rho = 0.83). (b) Corelation obtained for Nanopore with the cDNA protocol (Spearman's rho = 0.69). (c) Corelation obtained for Nanopore with the RNA direct protocol (Spearman's rho = 0.90).

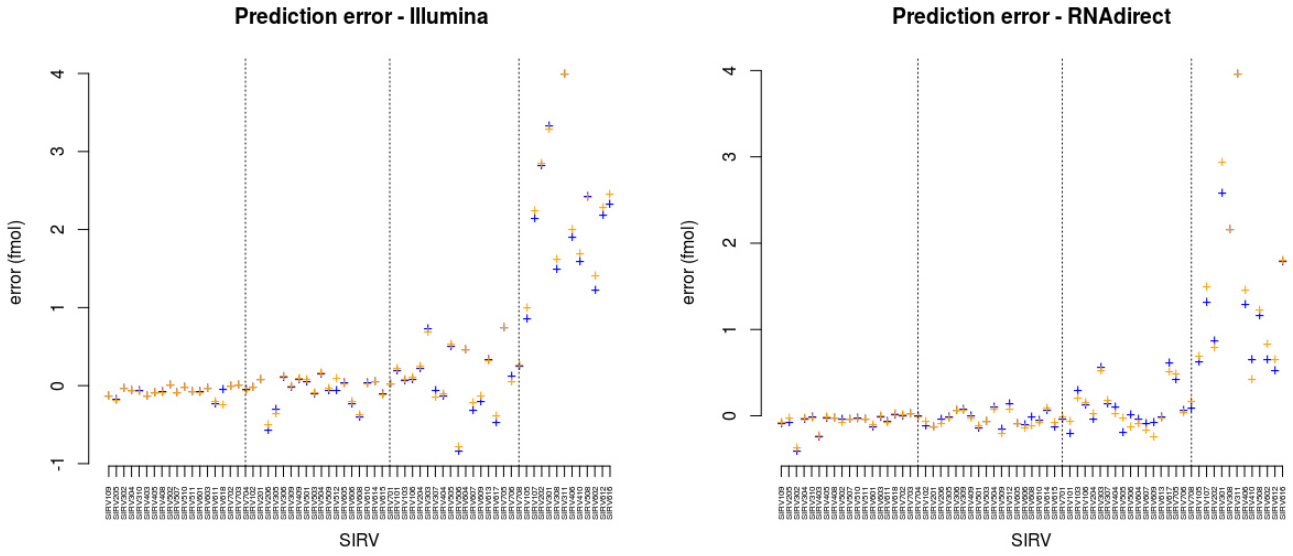


Figure 5: **Reproductibility of the prediction error.** The error between the prediction and the real quantification has bee computed for each replicates C2R1 and C2R2 for (a) the illumina dataset and (b) the Nanopore RNAdirect dataset.

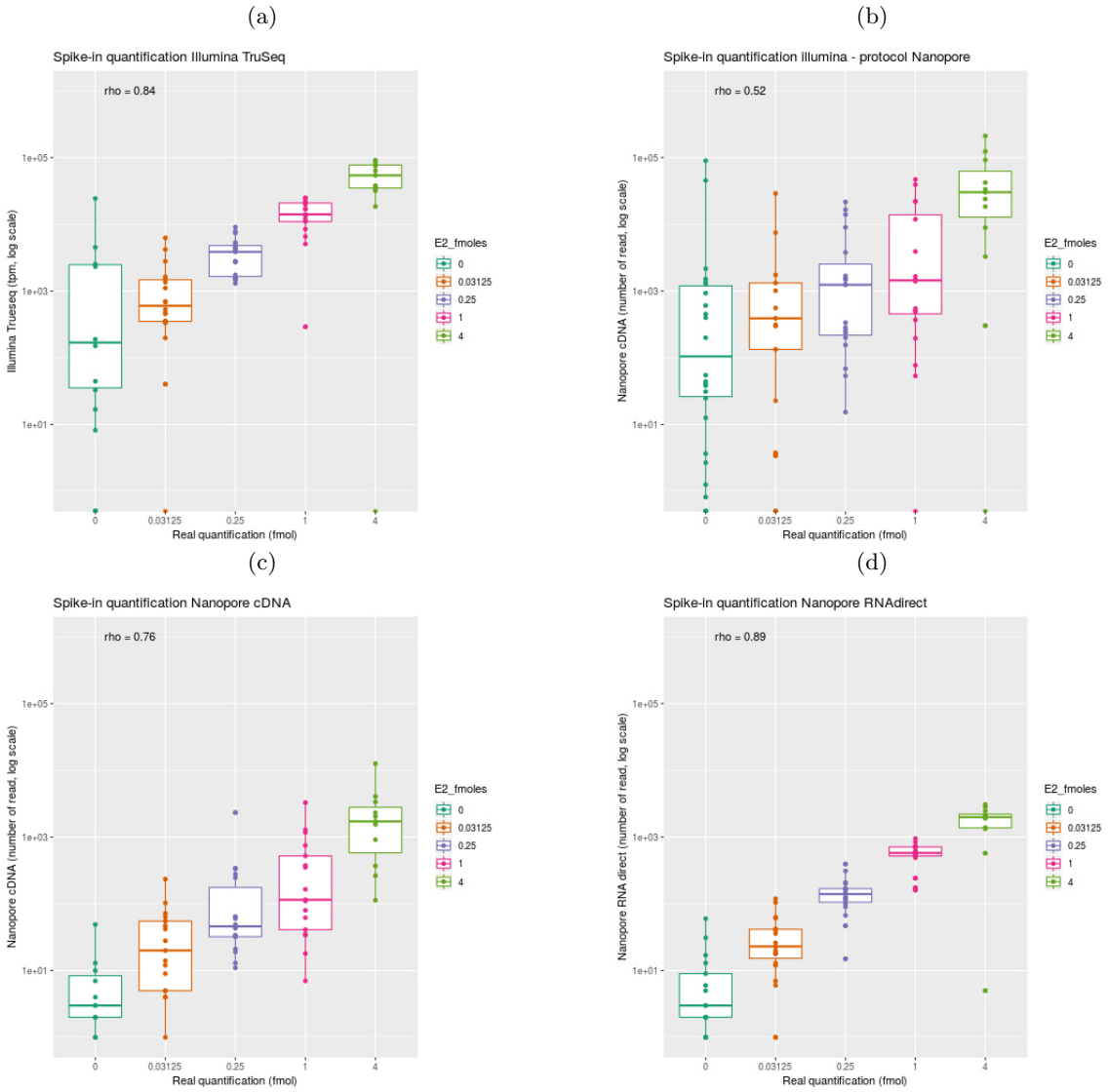


Figure 6: **Evaluation of quantification using the SIRV E2 spike-in mix and the over-annotation supplied by Lexogen** Reads were mapped against the SIRV transcriptome and quantifications computed at transcript level. The observed quantification are correlated with the known theoretical quantification of the spike in. (a) Correlation obtained for Illumina with the TruSeq protocol (Spearman's rho = 0.84). (b) Correlation obtained for illumina with the cDNA synthesis Nanopore protocol (Spearman's rho =0.52). (c) Corelation obtained for Nanopore with the cDNA protocol (Spearman's rho = 0.76). (d) Corelation obtained for Nanopore with the RNA direct protocol (Spearman's rho = 0.89).

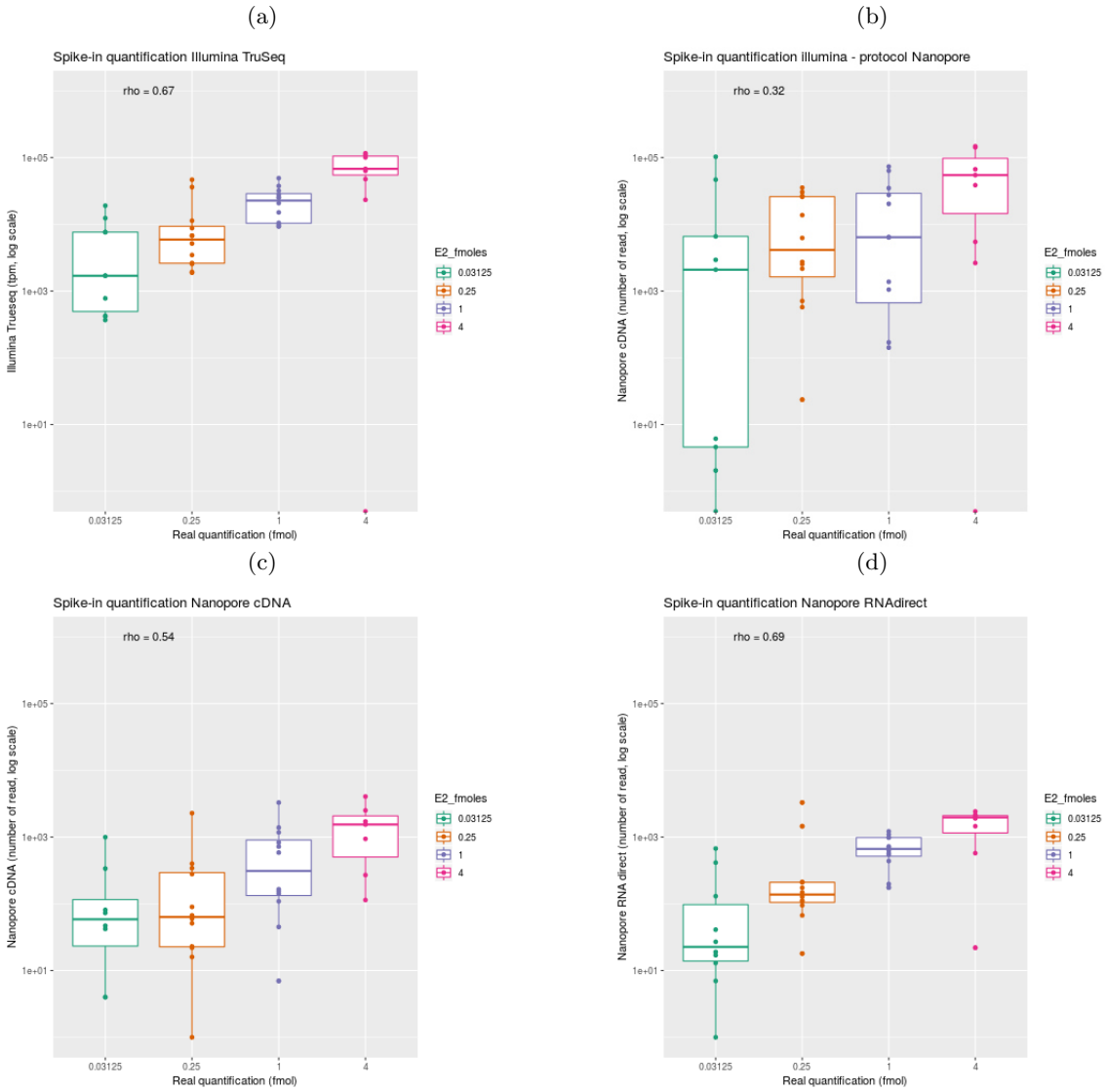
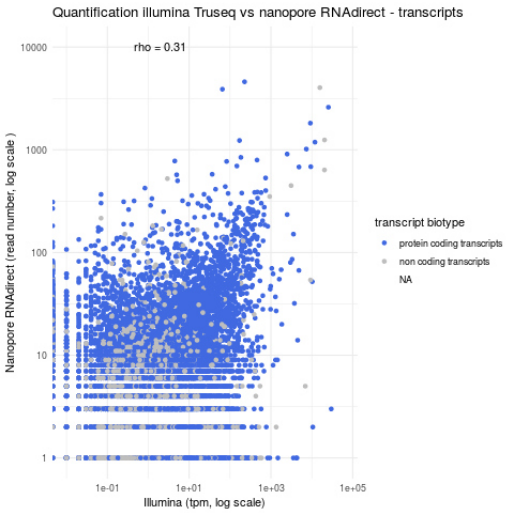
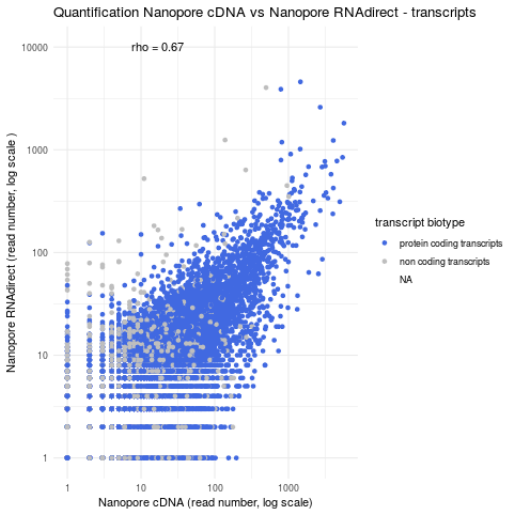


Figure 7: **Evaluation of quantification using the SIRV E2 spike-in mix and the incomplete annotation supplied by Lexogen** Reads were mapped against the SIRV transcriptome and quantifications computed at transcript level. The observed quantification are correlated with the known theoretical quantification of the spike in. (a) Correlation obtained for Illumina with the TruSeq protocol (Spearman's rho = 0.67). (b) Correlation obtained for illumina with the cDNA synthesis Nanopore protocol (Spearman's rho =0.32). (c) Corelation obtained for Nanopore with the cDNA protocol (Spearman's rho = 0.54). (d) Corelation obtained for Nanopore with the RNA direct protocol (Spearman's rho = 0.69).

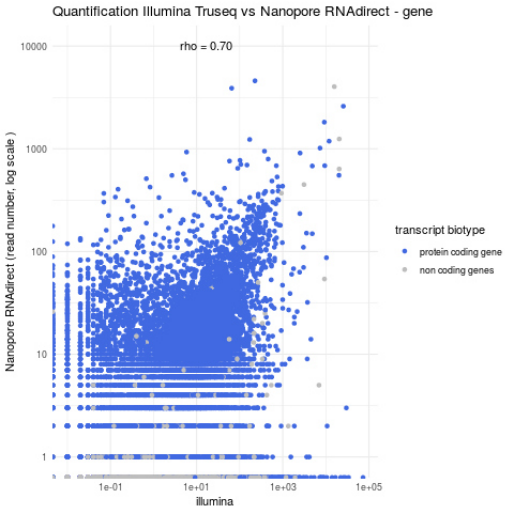
(a)



(b)



(c)



(d)

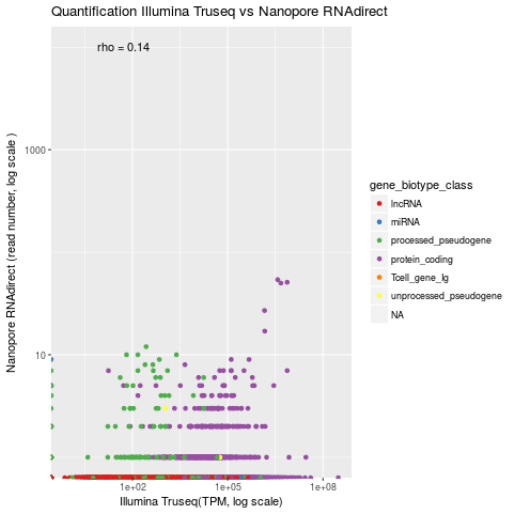


Figure 8: **Comparison of quantifications in Liver** Reads were mapped against the mouse reference transcriptome. Transcript annotated as coding protein transcript are in blue. Spearman's ρ has been computed for all transcripts. (a) Comparison of Nanopore RNA direct and Illumina (TruSeq) quantifications (Spearman's $\rho = 0.31$). (b) Comparison of Nanopore RNA direct and Nanopore cDNA quantifications (Spearman's $\rho = 0.67$). (c) Comparison of Nanopore RNA direct and Illumina (TruSeq) quantifications. Transcript quantification were summed for each gene. (Spearman's $\rho = 0.70$). (d) Reads were mapped against the mouse reference genome and quantifications computed at gene level. We compared the Nanopore RNAdirect and the illumina Truseq protocols (Spearman's $\rho = 0.14$). Green points correspond to processed pseudogenes.

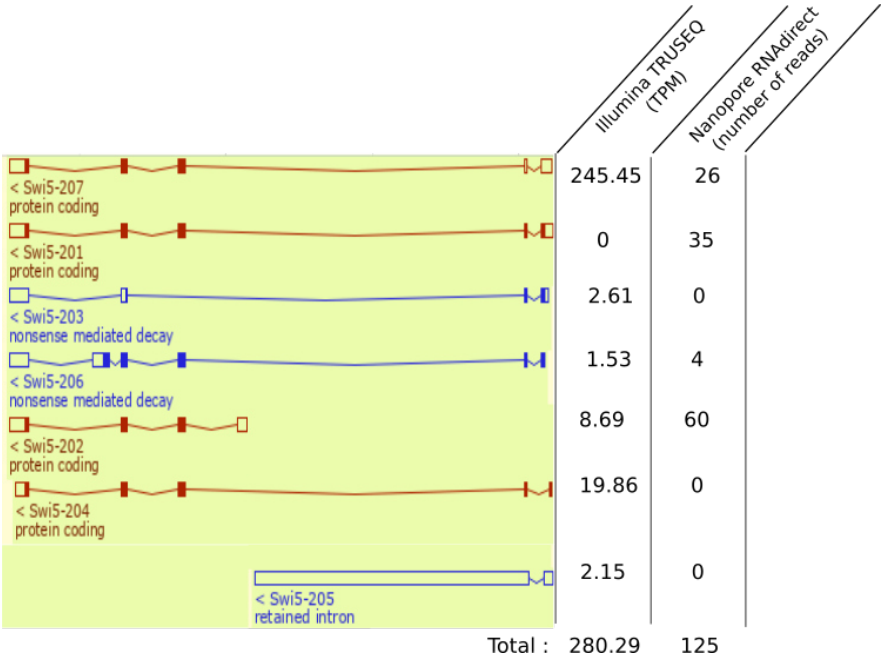
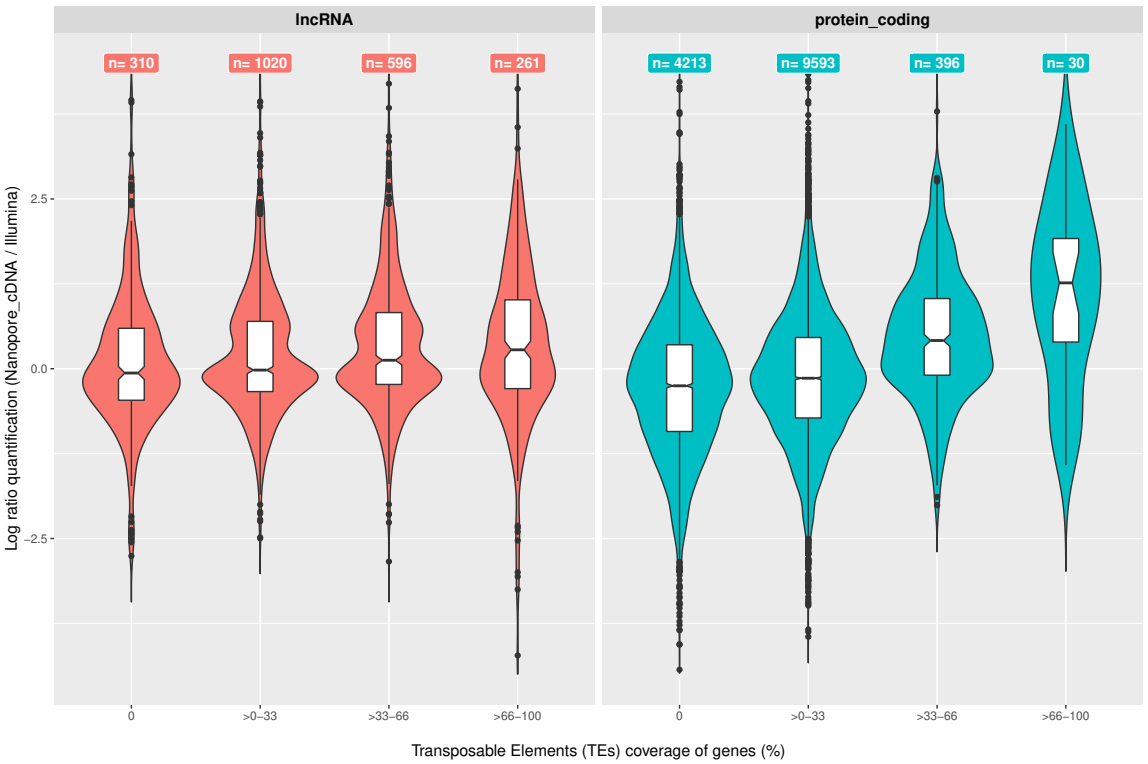


Figure 9: **Quantifications for Swi5 trancripts.** Swi5 annotation visualized with the Ensembl genome browser. The transcript Swi5-201 has no short read which uniquely maps to it. Therefore RSEM cannot allocate read to this transcript. With ONT RNA-Seq we have long enough reads to distinguish it from the other transcripts.

(a)



(b)

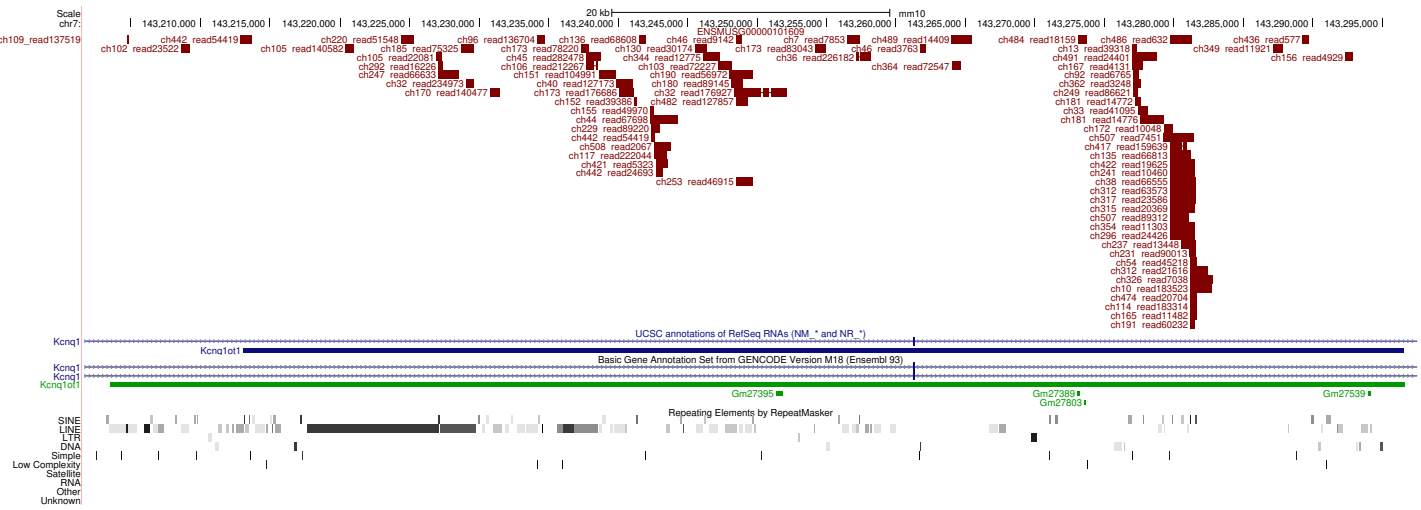
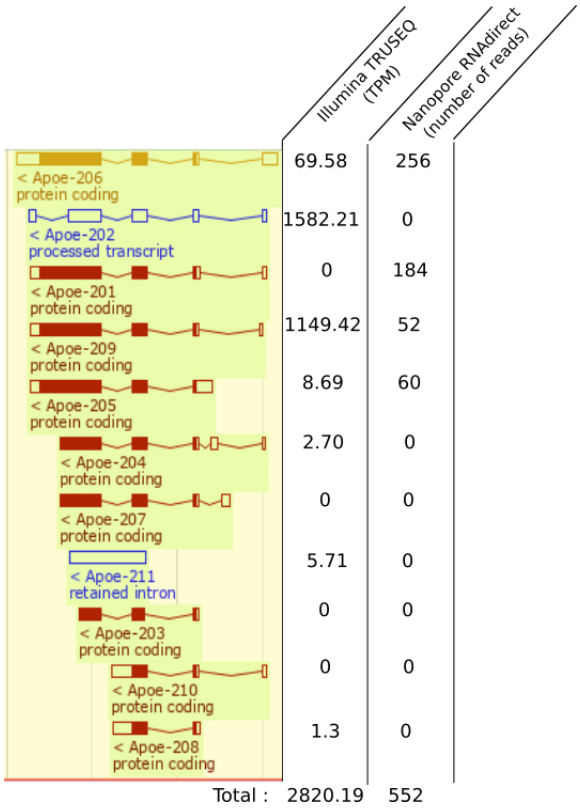


Figure 10: **Mouse lncRNA quantification in the ONT and Illumina cDNA-Seq.** (a) Gene quantification ratio (ONT cDNA-Seq versus Illumina cDNA-Seq) with variable gene coverage in Transposable Elements (TEs). Long non-coding RNAs (lncRNAs) are represented on the left panel (in red) while protein-coding genes are represented on the right (in blue). Only genes expressed in both conditions are represented (e.g. quantifications with $TPM > 0$ and $Nanopore_cDNA >= 1$). (b) UCSC screenshot of the KCNQ1OT1 locus. ONT cDNA-Seq reads mapped onto mm10 are represented on the top track, followed UCSC and GENCODE gene annotation tracks. Finally, the bottom tracks represents the repeat elements including transposable elements annotated by Repeat Masker.

(a)



(b)

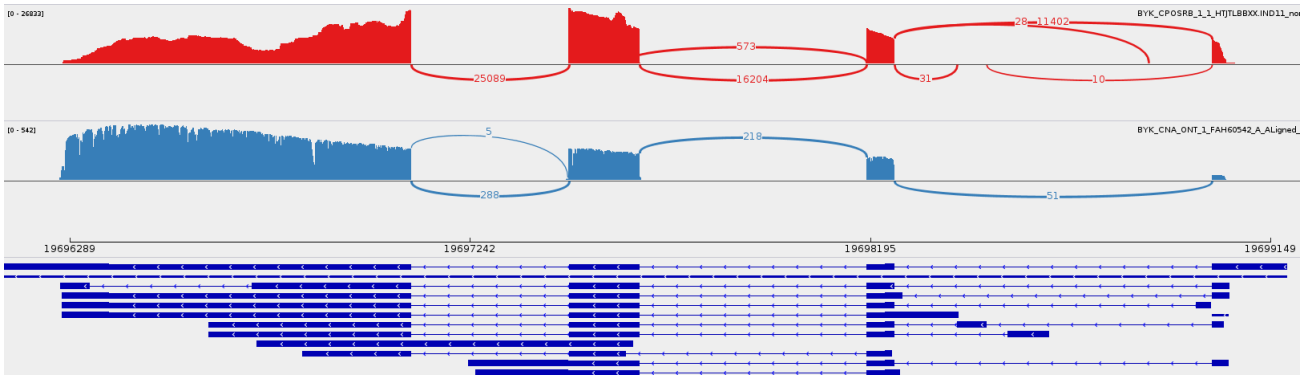


Figure 11: **Quantifications for APOE trancripts** (a) APOE annotation visualized with the Ensembl genome browser and quantification obtained with Illumina Truseq and RNAdirect (b) Sashimi plot obtained with IGV. Junctions covered by less than 5 reads were filtered out. First track shown Illumina Truseq reads and second track Nanopore RNA direct reads. As shown by the Sashimi plot, the most expressed transcript is not annotated. It Correspond to the transcript Apoe-206 with a shorter UTR.

4 | Perspectives

4.1 Résolution des problèmes de mapping multiple sur les transcrits

L'article présenté ci-dessus a, entre autre, permis de montrer que les longues lectures ne couvrent pas toutes des transcrits complets. Les lectures étant séquencées depuis la queue poly A (c'est-à-dire l'extrémité 3' des transcrits), les lectures sont tronquées plus ou moins loin au niveau de l'extrémité 5'. On observe un "effet escalier" dans les alignements : plus on avance vers l'extrémité 5' du gène et plus la couverture du gène diminue. La figure 17 illustre ce constat pour l'exemple du SIRV 505

De plus ces analyses ont aussi permis d'observer qu'il était encore difficile d'attribuer les lectures au bon transcrit d'un même gène. L'incomplétude des lectures en 5' explique en partie ce constat et il est possible d'utiliser le fait que les lectures soient complètes en 3' afin de résoudre certains problèmes de mapping multiple. En effet, habituellement, si une lecture s'aligne sur deux transcrits d'un même gène et que les scores associés aux deux alignements sont égaux, la lecture est attribuée à l'un ou l'autre des transcrits de façon aléatoire. Dans ce cas, on fait l'hypothèse que tous les transcrits ont la même abondance. En pratique, cette hypothèse est rarement vérifiée. C'est dans cette situation que nous proposons une amélioration de ce qui est fait actuellement pour résoudre les problèmes de mapping multiple.

FIGURE 17 – Couverture du SIRV 505 (RNAdirect). On observe ici l'effet escalier, plus l'on va vers l'extrémité 5' du transcrit et moins le SIRV est couvert.

Ce travail s'inscrit dans la continuité des travaux publiés dans le cadre de ma thèse mais a été repris et réalisé par Éric Cumunel, ingénieur dans l'équipe. J'ai aussi pu participer activement aux discussions liées à ce projet à mon retour de congé maternité. Je présente ici brièvement ses résultats. Ceux-ci permettent de mettre en évidence une application possible des résultats présentés ci-dessus.

La figure 18 présente le cas d'une lecture qui s'aligne avec un score identique sur les transcrits T1 et T2. Considérant que les lectures sont complètes en 3' et comme la lecture s'aligne sur T2 depuis l'extrémité 3' du transcrit, la méthode proposée ici permet d'attribuer la lecture à T2 plutôt qu'à T1. Dans ce cas particulier, on peut résoudre le problème de mapping multiple.

La limite principale de cette approche est l'incomplétude des annotations en 3'. Effectivement si les transcrits sont mal annotés en 3' alors il est complexe d'obtenir des résultats satisfaisants. Il se trouve que c'est régulièrement le cas, les bornes des UTR n'étant pas toujours bien définies dans les annotations.

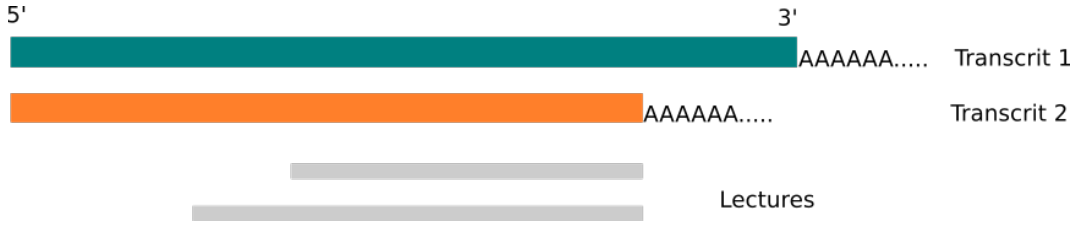


FIGURE 18 – Cas où l'on peut réattribuer une lecture à un nouveau transcrit. Les lectures représentées en gris s'alignent avec un score identique sur les deux transcrits. Cependant, leurs extrémités s'alignent parfaitement avec l'extrémité 3' du transcrit 2. On peut donc les attribuer au transcrit 2 plutôt qu'au transcrit 1.

Éric Cumunel à développé une méthode permettant d'automatiser la ré-attribution des lectures aux bons transcrits dans le cas présenté ci-dessus. Il a pu tester sa méthode sur les données de Spike-in. Les corrélations entre les quantifications réelles et les quantifications observées s'en trouvent améliorées. La valeur du rho de spearman passe de 0.86 à 0.95 par exemple pour le jeu de donnée RNA direct. Chez la souris, lorsque l'on compare les quantifications obtenues en RNAdirect et celles obtenues avec illumina, les corrélations augmentent après la ré-attribution des lectures aux transcrits de chacun des gènes : on passe de $\rho = 0.83$ à $\rho = 0.88$. En tout, 7.8% des lectures ont été ré-attribuées à un nouveau transcrit.

4.2 Évolution de la profondeur et des taux d'erreurs

De nouveaux jeux de données ont été séquencés afin de tester de nouveaux protocoles de préparation de librairies plus récents (protocole SQKPCS109 au lieu de SGKPCS108 pour le cDNA et SQK-RNA002 au lieu de SQK-RNA001 pour le RNAdirect). Ce jeu de données n'a pas été publié mais a aussi permis de tester cette nouvelle méthode. Les technologies de troisième génération évoluent très vite et le séquençage de ce nouveau jeu de données à permis d'augmenter les profondeurs obtenues précédemment. Pour le RNA direct, par exemple, nous avions, dans le jeu de données publié, 571 000 lectures dans le réplicat le plus

couvert alors qu'il y en a environ 5 millions dans ce nouveau jeu de données. Aussi, les taux d'erreurs ont baissé : nous sommes passés de plus de 14% à environ 12% pour le RNAdirect par exemple.

L'utilisation d'une nouvelle version du basecaller (c'est à dire d'une nouvelle version du logiciel qui permet de transformer le signal électrique obtenu au moment du séquençage en séquence nucléotidique), notée "high accuracy" dans la Table 1, a permis de faire baisser de nouveau ces taux d'erreurs. Ceux-ci sont présentés dans la table ci-dessous. Le jeu de données publié à lui aussi été de nouveau basecallé et cela a permis une diminution nette des taux d'erreurs. Il est intéressant de noter que l'utilisation de ce basecaller a permis de diviser par deux le taux d'erreur du jeu de donnée RNAseq qui a été publié. Une limite actuelle est que son utilisation est bien plus coûteuse en temps de calcul.

	basecaller classique	basecaller "high accuracy"
cDNA [SGKPCS108 - R 9.4.1]	11.9%	8.1%
cDNA [SQKPCS109 - R 9.4.1]	10.3%	8%
RNAdirect [SQK-RNA001 - R 9.4.1]	14.8%	7.5%
RNAdirect [SQK-RNA001 - R9.4.1]	11.9%	10%

TABLE 1 – Taux d'erreur obtenus pour les différents jeux de données. Les protocoles ainsi que la version du pore utilisés sont indiqués entre crochet.

Quatrième partie

Modélisation des évènements d'épissage complexes

1 | Background

Chez les organismes eucaryotes, les gènes contiennent des exons et des introns. Lors de la maturation des molécules de pré-ARN messager et plus particulièrement lors de l'épissage, les introns sont retirés. Plusieurs choix d'appariement de sites d'épissage entre eux sont possibles et un seul gène peut alors produire plusieurs transcrits matures différents et donc parfois plusieurs protéines différentes. C'est ce que l'on appelle l'épissage alternatif (cf introduction de la thèse, paragraphe 2.3).

1.1 Méthodes bio-informatiques existantes

De nombreuses méthodes pour étudier l'épissage alternatif depuis des données RNAseq existent. La plupart sont basées sur l'alignement des lectures courtes sur un génome de référence et/ou sur les annotations. Elles peuvent être globales, c'est à dire, basées sur la reconstruction des transcrits complets comme Stringtie [58], Cufflinks[76] ou FlipFlop [11] ou bien locales (exon-centrées) comme MISO [29], DEXseq [3] ou encore AS-Quant[16]. Il est aussi possible d'étudier l'épissage sans génome de référence, en assemblant les transcrits complets par exemple avec Trinity[22] ou bien Oases[70]. Enfin, KisSplice[64] est un assembleur local de transcriptome et permet, en assemblant les lectures courtes d'étudier notamment les événements d'épissage à l'échelle de l'exon.

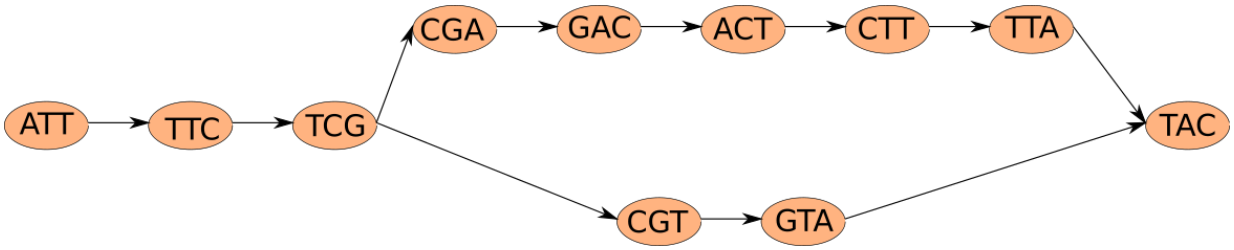


FIGURE 19 – Exemple d'un graphe de De Bruijn construit à partir de deux séquences : ATTCGACTTAC et ATTCGTAC avec k=3.

1.2 KisSplice et Graphes de De Bruijn

Comme beaucoup d'assembleurs, dans un premier temps, KisSplice reconstruit le graphe de De Bruijn à partir des lectures. Celui-ci nous permet de détecter des évenements d'épissage simple. On parle d'évènements d'épissage simple lorsque l'on compare seulement deux transcrits alternatifs entre eux.

1.2.1 Construction du graphe de De Bruijn (DBG)

Chaque lecture est coupée en mots de taille k, nommés k-mers. Chaque k-mer constitue un noeud du graphe et les noeuds sont reliés entre eux par une arête si ils partagent (k-1) nucléotides et que le suffixe du premier noeud est égale au préfixe du second noeud. La figure 19 présente un exemple de graphe de De Bruijn construit à partir de deux exemples de lectures de taille 8 et avec k=3. Dans la pratique, les lectures ont la plupart du temps une taille comprise entre 100 et 200 paires de base. Dans le travail présenté ici, les valeurs de k utilisées sont comprises entre 31 et 41 nt.

1.2.2 Détection des événements d'épissage

Dans un graphe de De Bruijn, un événement d'épissage simple forme une structure particulière appelée bulle. Une bulle est une paire de chemins partageant uniquement deux noeuds : le noeud source et le noeud cible. Un exemple de bulle est représenté figure 19. Les

extrémités des exons flanquants correspondent aux noeuds "TCG" et "GAG" sur la figure 19. Les deux chemins entre ces deux noeuds correspondent aux deux variations d'épissage observées. La partie génomique dite variable (ici "ACT") est incluse dans le chemin du haut sur la figure 19 et exclue dans le chemin du bas. Celui-ci correspond à la jonction entre les deux exons flanquants. Un algorithme d'énumération de bulles dans le graphe nous permet ensuite de détecter des évènements d'épissage simple. Les bulles doivent remplir deux conditions :

— Le chemin d'exclusion - c'est à dire le chemin le plus court- doit avoir une taille inférieure à 2k-2 nucléotides. Cette taille correspond à la longueur de la séquence correspondant à la jonction entre les deux exons flanquants ce qui permet de différentier les bulles dues à un évènement d'épissage alternatif de celles dues à des SNPs (single nucleotide polymorphism). Ces dernières ayant des chemin de tailles 2k-1.

— Chacun des chemins doit compter moins de cinq noeuds branchants.

On définit un noeud branchant comme un noeud de degré supérieur ou égal à deux (c'est-à-dire ayant au moins deux arêtes sortantes ou deux arêtes entrantes). Cette dernière condition permet d'éviter d'énumérer des bulles présentes dans des régions du graphe correspondant à des répétitions [42]. Ces répétitions proviennent le plus souvent d'éléments transposables comme l'élément Alu très présent dans le génome humain par exemple. Elles forment des structures particulières dans le graphe et lient entre eux des gènes différents que l'on aurait voulu pouvoir distinguer les uns des autres. Ces régions du graphe sont très branchantes et limiter le nombre de noeuds branchants autorisés dans les bulles permet de ne pas les explorer.

1.3 Évènements d'épissage complexe

On définit un événement d'épissage complexe comme l'ensemble des variations observées entre deux sites d'épissage constitutifs à tous les transcrits d'un même gène. Un événement d'épissage complexe permet donc de produire plus de deux transcrits alternatifs.

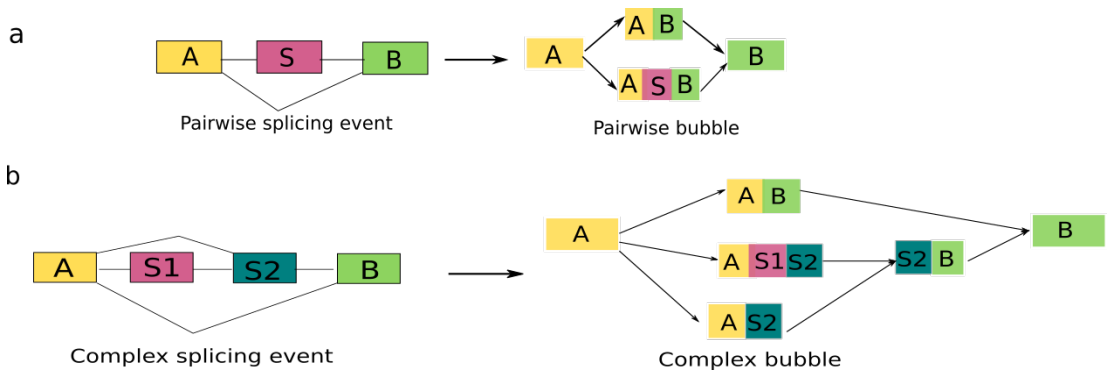


FIGURE 20 – les événements d'épissage forment des structures dans le DBG(a) Exemple d'un graphe d'épissage pour un saut d'exon (=événement d'épissage simple) et la bulle correspondante dans le DBG. (b) Exemple d'un évènement d'épissage complexe et de la bulle complexe correspondante.

La figure 20a présente un exemple d'évènement d'épissage simple (le saut de l'exon S) et la figure 20b présente un événement d'épissage complexe. Celui-ci correspond à la combinaison du saut de l'exon S1 et du double saut des exons S1 et S2 . Dans le graphe de De Bruijn, un évènement d'épissage complexe correspond à une bulle complexe. Les bulles simples sont en fait un cas particulier de bulles complexes. On estime que 30% des évènements d'épissage sont complexes [72].

Quelques méthodes comme MAJIQ[79], Whippet [72], Leafcutter[40], FRASER [53] et McSplicer[2] permettent d'étudier les évènements d'épissage complexe. Elles sont toutes basées sur un génome de référence ou sur des annotations. On se propose ici d'étudier l'épissage complexe sans génome de référence ni annotation. Cela permet de détecter et quantifier les événements d'épissage complexe chez les espèces non modèles mais aussi de détecter de nouveaux sites d'épissage (non annotés) chez les espèces modèles. De plus, même dans le cas d'espèces pour lesquelles un génome de référence est bien connu, les approches de novo (c'est-à-dire sans génome de référence) d'étude de l'épissage alternatif, permettent de détecter des évènements non vus par les méthodes basées sur un génome de référence [8].

1.4 Régulation des évènements d'épissage complexe

On sait que toutes les variations d'épissage observées ne sont pas régulées. [77][59]. On peut alors chercher à identifier celles qui le sont. L'observation de quantifications significativement différentes entre deux conditions expérimentales est un argument pour soutenir l'idée qu'un événement d'épissage est régulé. Plus précisément, on cherche à identifier des régions génomiques différentiellement incluses entre les deux conditions expérimentales étudiées. Ainsi, j'ai développé une méthode d'analyse différentielle pour les évènements complexes basée sur la généralisation du modèle de kissDE [45] à la comparaison de plus de deux transcrits à la fois. On étudiera ici des exemples d'événements d'épissage provenant de deux jeux de données déjà publiés. Le premier basé sur la déplétion d'un facteur d'épissage déjà bien connu : PTBP1 [44]. Le second permet d'étudier l'impact de la déplétion d'un deuxième facteur d'épissage, RED, dont les modes d'action dans la cellule sont moins bien connus [5]. Ce jeu de données permet d'étudier l'impact de l'infection des cellules humaines par le virus de la grippe (IAV) mais aussi l'impact de la déplétion de RED sur des cellules infectées ou non. On se focalise ici sur les conditions expérimentales permettant d'étudier l'impact du knock-down de RED sur des cellules saines.

2 | Pipeline d'analyse des événements complexes

2.1 Présentation générale

Pour identifier les cas où le modèle pairwise de KisSplice n'est pas suffisant pour expliquer les variations d'épissage observées, j'ai mis en place un pipeline d'identification et de quantification des événements d'épissage complexe (figure 21). Celui-ci est un post-traitement de KisSplice et permet de regrouper entre eux les événements pairwise qui le peuvent au sein d'événements d'épissage complexe.

On utilise le fait que les bulles sorties par KisSplice ont au maximum 5 noeuds branchants par chemin pour éviter de considérer des zones plus complexes du graphe de De Bruijn correspondant à des répétitions intergéniques. Cela nous permet d'éviter de construire des bulles complexes liant différents gènes distincts les uns aux autres. Ensuite, les bulles complexes sont construites puis quantifiées.

2.2 Définition des bulles complexes

Dans le graphe de De Bruijn, les événements d'épissage complexe correspondent à des bulles complexes. On peut définir une bulle complexe de la manière suivante. Une (s,c)-bulle complexe B, dans un graphe dirigé G, est un sous graphe de G tel que :

1. B est le sous-graphe induit par AP, AP étant l'ensemble de tous les (s,c)-chemins dans

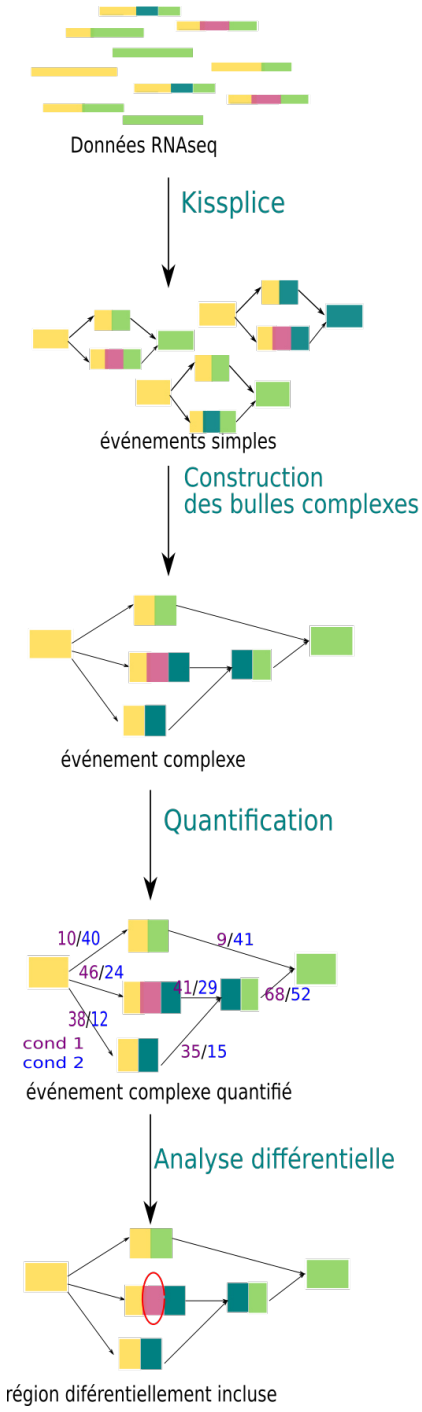


FIGURE 21 – Pipeline d'identification et de quantification des événements d'épissage complexe. Les lectures sont assemblées avec KisSplice et les événements simples identifiés. Ce sont des bulles dans le DBG. Les bulles complexes sont ensuite construites depuis le graphe induit par les bulles simples. Elles sont ensuite quantifiées. Enfin les régions génomiques différentiellement incluses entre les conditions expérimentales testées sont identifiées (analyse différentielle)

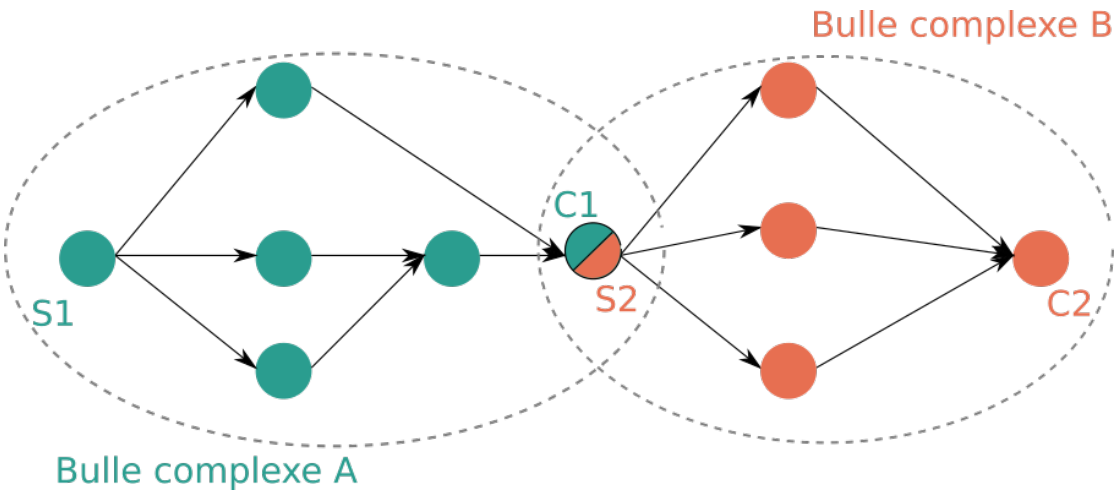


FIGURE 22 – La condition 4 de la définition ci-dessus permet de séparer les bulles complexes A et B en deux bulles complexes distinctes

G.

2. s est la seule source de B (seul le noeud s permet de rejoindre toutes les arêtes $v \in B$)
3. c est la seule cible de B (seul le noeud t peut être rejoint par toutes les arêtes $v \in B$)
4. s et c sont les seuls noeuds communs à tous les chemins $p \in AP$

Une (s,c)-bulle complexe maximale est une (s,c)-bulle complexe non incluse dans une autre bulle complexe. Pour la suite nous parlerons toujours de bulles complexes maximales.

Du point de vue de l'implémentation, la condition 4 énoncée ci-dessus n'est pas prise en compte directement. Celle-ci permet de séparer des bulles qui seraient reliées entre elles par un noeud ou une partie linéaire du graphe de De Bruijn. Cela est illustré figure 22 où la condition 4 permet de séparer les bulles complexes A et B. Étant donné que l'on se base sur le graphe induit par les bulles déjà énumérées par KisSplice, on ne garde que les k-mers déjà inclus dans une bulle simple (on n'utilise pas l'option –output-context de KisSplice qui nous donnerait les séquences des unitigs flanquants de la bulle plutôt que les k-mers) et les bulles sont donc déjà séparées les unes des autres dans le cas où la longueur de l'unitig C1 est supérieur à k. En pratique, cela arrive rarement, les unitigs sont souvent bien plus longs que k.

Cette définition diffère de celle donnée des superbulles dans la littérature [56] en un point principal : On autorise ici la présence de noeuds branchants au sein de chacun des chemins des bulles complexes. En effet, dans la littérature les superbulles sont définies comme l'ensemble des noeuds atteignables depuis un noeud source ainsi que depuis un noeud cible. Cette définition ne permet pas aux chemins de la bulle de contenir des noeud branchants. Cette différence nous permet d'énumérer des évènements d'épissage complexe qui recouvrent des régions répétées. Cela semble pertinent, par exemple chez l'Homme, ou la présence d'éléments répétés (de type Alu) exonisés est fréquente.

2.3 Construction des bulles complexes

La reconstruction des bulles complexes implique l'identification de paires de noeuds sources et cibles potentielles des bulles complexes. Pour cela, on utilise les noeuds déjà considérés comme source ou cible des bulles simples sorties par KisSplice. Toutes les paires possibles sont considérées.

Ensuite, pour chaque paire de noeuds source (s) /cible (c), on identifie tous les noeuds atteignables depuis s et c. Pour cela on utilise un algorithme de parcours en largeur du graphe. On considère une profondeur limite (max_depth) à partir de laquelle on s'arrête d'explorer le graphe. On construit ainsi deux ensembles de noeud : celui des noeuds atteignables depuis s et celui des noeuds atteignables depuis c. On cherche ensuite l'intersection I de ces deux ensembles de noeuds. L'ensemble de noeuds I correspondant forme une bulle complexe s'il remplit les conditions suivantes :

1. I n'est pas vide.
2. s appartient à I.
3. c appartient à I.

De plus, étant donné les étapes précédentes (construction de I grâce à deux BFS), s est la seule source et c est la seule cible de chacune des bulles complexes ainsi reconstruites.

2.4 Modélisation du choix des sites d'épissage

Afin de sélectionner les arêtes à quantifier pour chacun des noeuds branchants il est nécessaire d'identifier les sites d'épissage atteignables depuis le noeud étudié. En effet, le choix des sites d'épissage ne correspond pas toujours aux arêtes présentes dans le graphe de De Bruijn (DBG). Étant donné que l'ARN n'est composé que de 4 bases différentes, il y a au maximum 4 arêtes sortantes ou entrantes d'un noeud dans le DBG. De même, il n'est pas rare que deux exons, appartenant à l'évènement d'épissage complexe considéré, commencent par le même nucléotide. Ainsi, depuis un noeud branchant, si l'on peut rejoindre deux exons commençant par le même nucléotide, alors il n'y aura qu'une seule arête dans le DBG. Par exemple, figure 23a les exons rose et vert commencent tous deux par une cytosine, cela se traduit par la présence d'une seule arête sur la figure 23b (entre le T et le C). Il est nécessaire de différentier ces deux options et donc de considérer, au moment de l'analyse différentielle, deux arêtes distinctes. Ce cas particulier est illustré figure 23. Sur le graphe d'épissage présenté figure 23a, le site d'épissage donneur de l'exon jaune peut être associé à trois sites d'épissage différents : le site d'épissage accepteur de l'exon rose, celui de l'exon gris ou bien celui de l'exon vert. Or, les exons rose et vert commencent tous deux par une cytosine. Le DBG correspondant est représenté figure 23b et l'on peut voir que le noeud jaune à un degré sortant de 2. Depuis le k-mer correspondant à la fin de l'exon jaune on peut aller vers un C ou bien vers un A. Afin de dissocier les trois possibilités, on décide de quantifier les arêtes du DBG correspondant aux trois sites d'épissage possibles (figure 23c).

Pour cela, on réalise un parcours de graphe en largeur (BFS) du DBG induit par les bulles sorties par KisSplice, à partir du noeud dont l'on souhaite quantifier les arêtes avec un critère d'arrêt n. n étant le nombre maximal de noeuds branchants atteignables, depuis le noeud étudié, dans l'arbre T, en sortie du BFS. Par défaut, on prend $n = 2$. On considère qu'il est nécessaire de sélectionner des arêtes différentes des arêtes sortantes du noeud racine si les noeuds vers lesquels celles-ci pointent sont eux-même branchants. En effet s'ils n'étaient pas branchants, on rentrerait dans une partie linéaire du graphe correspondant à la partie

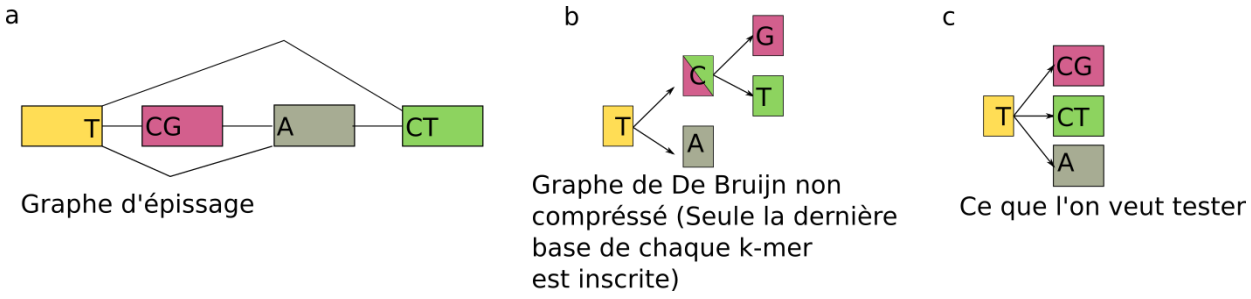


FIGURE 23 – Modélisation du choix des sites d'épissage : sélection des arêtes à quantifier dans le graphe de De Bruijn. Chaque arête choisie correspond à une option possible pour le spliceosome afin d'apparier le site d'épissage donneur de l'exon jaune avec les sites d'épissage accepteurs possibles.

variable étudiée (par exemple un exon skippé). Il n'y aurait donc pas de possibilités de choix supplémentaire pour le spliceosome à cet endroit-là. On répète ce raisonnement n fois. On parcourt alors l'arbre T depuis les feuilles jusqu'à la racine. Pour chacune des feuilles f de T, si le noeud parent de f est de degré sortant supérieur à deux, on liste les arêtes sortantes de ce noeud. Puis on vérifie la séquence de chacune de ces arêtes afin de sélectionner celle qui mène effectivement au noeud f considéré. On répète ensuite l'opération pour chacun des niveaux de l'arbre T depuis les feuilles, jusqu'à la racine. Les noeuds déjà considérés sont marqués comme vu afin de ne pas les considérer une deuxième fois par la suite.

2.5 Quantification des bulles complexes

Ensuite, chacune des bulles complexes est quantifiée. On cherche à quantifier les arêtes sortantes et entrantes de chacun des noeuds branchants du DBG. Pour cela, on s'appuie sur le fait qu'une arête du graphe de De Bruijn correspond à un (k+1)-mer. On cherche donc à quantifier les k+1-mers correspondants aux arêtes sélectionnées. Il n'existe pas à notre connaissance de librairie permettant de quantifier les arêtes d'un DBG mais on peut facilement quantifier les noeuds avec pyGATB. Ainsi, pour chaque réplicat, on reconstruit avec pyGATB le graphe de De Bruijn des (k+1)-mers. On cherche ensuite pour chaque

arête à quantifier (appartenant au DBG des k-mers) le noeud correspondant dans le graphe des (k+1)-mers. On peut ainsi obtenir, pour chaque arête précédemment sélectionnée, les quantifications pour chacun des réplicats considérés dans l'analyse.

2.6 Analyse différentielle

Enfin, on réalise une analyse différentielle afin de détecter les variations d'épissage régulées entre les conditions expérimentales étudiées. On cherche à identifier des régions génomiques différentiellement incluses entre les conditions étudiées. On réalise ainsi plusieurs tests pour chaque évènement complexe. En effet toutes les variations observées au sein d'un événement ne sont pas forcément régulées entre les conditions étudiées.

Le modèle statistique de kissDE [45] a été généralisé au cas où l'on compare plus de deux transcrits entre eux. À l'intérieur des bulles complexes, on considère que chaque noeud de degré entrant ou sortant supérieur à deux correspond à un choix du spliceosome. Pour chacun d'entre eux, on teste l'usage des arêtes c'est-à-dire l'usage des introns et donc le choix des sites d'épissage. On teste l'hypothèse nulle suivante : l'usage des sites d'épissage ne varie pas entre les conditions expérimentales étudiées.

Pour cela, les comptages sont normalisés en utilisant la normalisation proposée par le paquet DESeq2 [46] par défaut. Puis, les comptages sont modélisés par une distribution binomiale négative et le paramètre de sur-dispersion est calculé. Ensuite, on utilise un modèle linéaire généralisé afin de modéliser l'expression des différents transcrits.

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

On teste l'interaction entre l'effet d'un variant i et l'effet de la condition expérimentale j sur le niveau d'expression du transcrit considéré. Les deux modèles présentés ci-dessus étant

emboîtés, on peut réaliser un test de rapport de vraisemblance. L'hypothèse nulle de ce test est la suivante : $H0 : (\alpha\beta)_{ij} = 0$.

On calcule alors une p-valeur pour chaque noeud branchant testé. Réaliser de nombreux tests successifs augmente la probabilité de trouver des résultats significatifs par hasard. Ainsi, les p-valeurs sont corrigées pour prendre en compte ce biais avec la méthode de Benjamini-Hochberg [7].

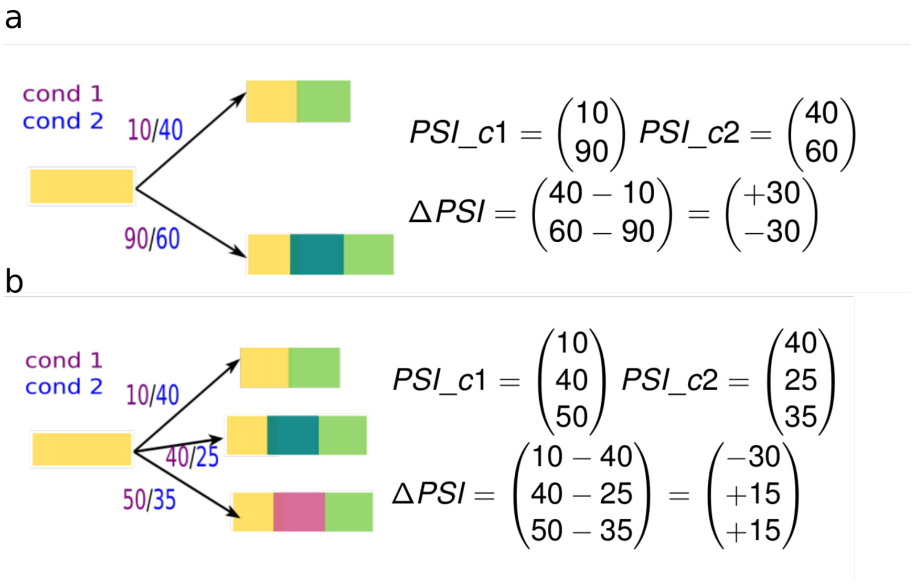


FIGURE 24 – Exemples de calcul de PSI et de ΔPSI (a) dans le cas d'un événement d'épissage simple et (b) dans le cas d'un événement d'épissage complexe où l'on compare trois transcrits. Les chiffres indiqués en couleur sur les arêtes correspondent au nombre de lectures supportant chacune des arêtes pour chaque condition expérimentale notée cond1 et cond2.

Enfin, on calcule la magnitude de l'effet observé grâce à la métrique PSI généralisée au cas où l'on compare plus de trois transcrits. Classiquement PSI (percent spliced in) reflète la proportion du transcrit d'inclusion c'est-à-dire le transcrit qui contient la région génomique dite variable (qui peut être incluse ou exclue de l'événement d'épissage). Ici, PSI devient un vecteur dans $\mathbb{R}^n$, n étant le nombre de transcrits comparés. Pour un test donné, chacune des valeurs de PSI traduit la proportion observée du transcrit correspondant parmi tous

les transcrits étudiés. La somme des ces valeurs est toujours égale à 1. On généralise aussi le calcul des ΔPSI, en soustrayant entre eux les vecteurs de PSI correspondant à chaque condition expérimentale étudiée. On obtient ainsi un vecteur de ΔPSI dont la somme des valeurs est toujours égale à 0. La figure 24 montre un exemple de calcul de PSI et de ΔPSI dans le cas classique avec deux transcrits (figure 24a) et généralisé à trois transcrits (figure 24b).

On ne souhaite pas réaliser les tests ni calculer les valeurs des ΔPSI pour les jonctions pour lesquels les comptages sont trop faibles, ainsi on filtre les cas pour lesquels la somme des comptages de chaque réplicat, quelle que soit la condition, est inférieure à 10. La valeur de ce seuil est discutée paragraphe 4.5.

3 | Pertinence de la modélisation

Pour tester la pertinence du modèle mis en place j'ai décidé d'étudier des exemples de gènes pour lesquels on peut formuler des hypothèses par rapport aux mécanismes sous-jacents aux variations d'épissage observées. Pour cela, j'ai utilisé deux jeux de données déjà publiés. L'un étudiant la déplétion du facteur d'épissage PTBP1 [44]. PTBP1 est un facteur d'épissage dont les cibles sont connues. Les mécanismes de régulation de l'épissage de certaines de ces cibles sont décrits dans la littérature [44] [80]. Le deuxième jeu de données concerne la dépletion d'un autre facteur d'épissage, RED, connu pour jouer un rôle important lors de l'infection des cellules humaines par le virus de la grippe [4][17]. Tous les mécanismes moléculaires utilisés par RED ne sont pas connus mais l'on sait par exemple que RED aide à réguler l'épissage des petits introns [30]. Ensuite, les résultats obtenus sont comparés à ceux de KisSplice [64] utilisé seul et avec ceux obtenus avec MAJIQ [79], une méthode d'analyse locale de l'épissage complexe largement utilisée.

3.1 Exemple d'évènements régulés par PTBP1

Tout d'abord, j'ai choisi les exemples de gène dont la régulation par PTBP1 est bien connue [44]. Dans leur publication, Linares et al [44] ont étudié l'effet de la déplétion de PTBP1 et PTBP2 sur différentes lignées cellulaires neurales différenciées ou non (ESC, NPC et MN). J'ai choisi d'étudier les données provenant de la lignée cellullaire NPC (Neural Progenitor Cell), c'est-à-dire de cellules neurales en cours de différenciation. De plus, des données iclip sont disponibles pour PTBP1. Ces données permettent d'identifier les points de

fixation de PTBP1 sur les transcrits des gènes étudiés. On peut alors connaître les points de fixation de PTBP1 sur le brin de pré-ARN messager. Cela nous permet de mieux comprendre les mécanismes moléculaires qui permettent de réguler les variations d'épissage observées et ainsi de pouvoir vérifier que nos résultats sont bien en accord avec les éléments connus d'un point de vue biologique. Enfin, des validations expérimentales permettent de confirmer ces observations [44].

3.1.1 Pbx1

Le gène Pbx1, et plus particulièrement le saut de l'exon 7, est connu pour être régulé par PTBP1 [44]. Les données publiées par les auteurs ont été téléchargées depuis SRA puis les lectures ont été ré-alignées sur le génome de référence de la souris (mm10) avec STAR (version 2.7.7). Enfin, les lectures alignées sur Pbx1 ont été sélectionnées.

La figure 25 présente le sashimi plot correspondant à ce saut d'exon. Chaque ligne correspond à un réplicat, les deux premières lignes correspondent à la condition contrôle et les deux dernières à la déplétion de PTBP1. La couverture, c'est-à-dire le nombre de lectures alignées pour chaque position génomique, est représentée en ordonnée. Ainsi, les pics de couverture correspondent aux exons. Les jonctions entre exons sont représentées par des arêtes entre les différents sites d'épissage et le nombre de lectures soutenant chacune de ces jonctions est indiqué. Le site de fixation de PTBP1, obtenu grâce aux données iclip, est reporté sur le sashimi plot avec une flèche rouge. On observe que l'exon 7 de Pbx1 est inclus plus souvent dans les transcrits de Pbx1 lorsque PTBP1 est déplété. Ce résultat est confirmé par la validation expérimentale réalisée par les auteurs et reproduite figure 25c. PTBP1 semble jouer un rôle dans la reconnaissance des sites d'épissage bornant l'exon 7. Quand PTBP1 est déplété alors l'exon est mieux reconnu. Les donnée iclip nous indiquent un point de fixation de PTBP1 en amont de l'exon 7 et confortent ces observations. On peut émettre l'hypothèse qu'en condition contrôle, c'est-à-dire lorsque PTBP1 est présent, alors les sites d'épissage de l'exon 7 sont plus souvent masqués par PTB. Ainsi l'exon serait moins bien reconnu et donc

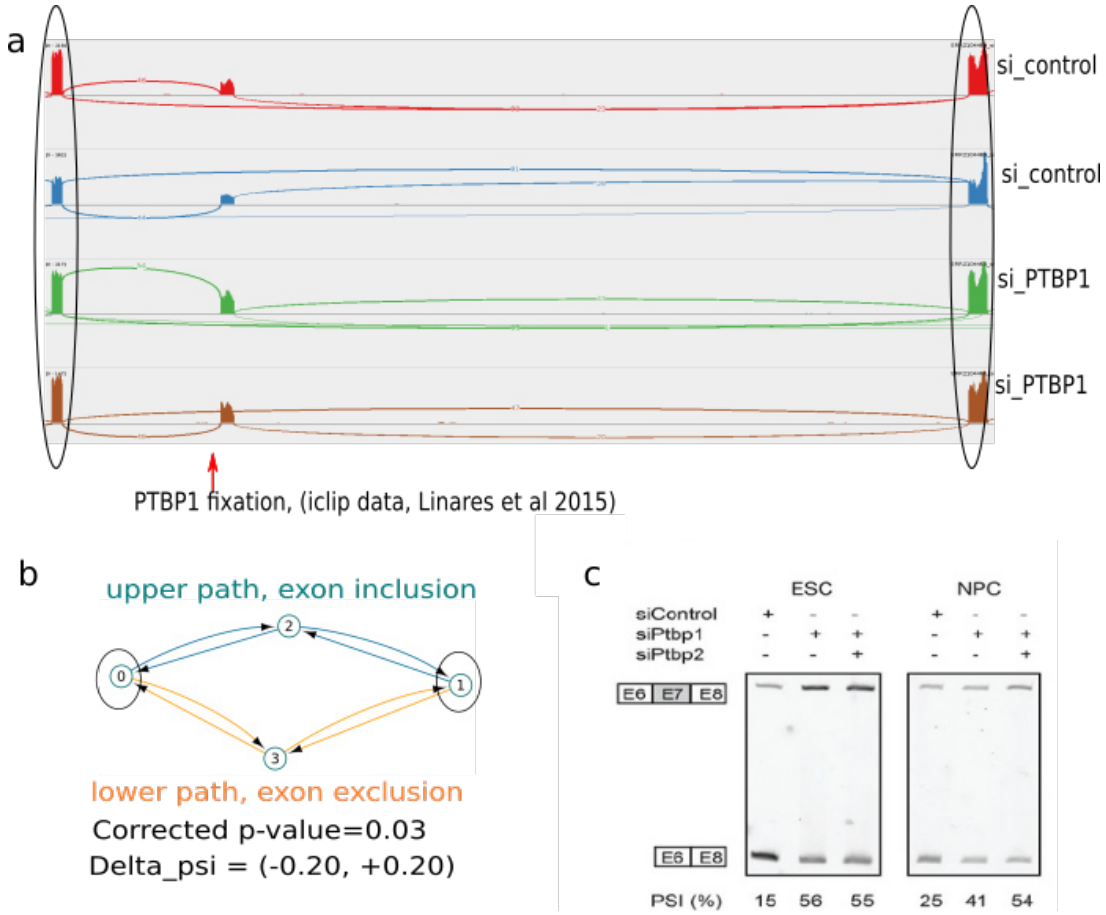


FIGURE 25 – Exemple du saut de l'exon 7 de Pbx1. (a) Sashimi plot correspondant au saut de l'exon 7. La couverture est représentée en ordonnée pour chacun des réplicats (2 par conditions). Les jonctions entre exons sont représentées par des arêtes et le nombre de lectures soutenant chacune de ces jonctions est indiqué sur les arêtes correspondantes (b) Bulle pairwise obtenue depuis le graphe de De Bruijn et correspondant au saut de l'exon 7 (c) Validation expérimentale par RT-PCR telle qu'elle apparait dans [44]. Le gel de droite correspond à la lignée cellulaire NPC décrite ici. La première colonne correspond à la condition contrôle, la deuxième colonne correspond à la déplétion de PTBP1

moins souvent inclus.

Ces résultats peuvent être confirmés sans génome de référence. La figure 25b représente la bulle reconstruite dans le graphe de De Bruijn correspondant au saut de l'exon 7 de Pbx1. On étudie ici les variations d'épissage observées entre les noeuds 0 et 3. Ceux-ci correspondent aux exons constitutifs autour de l'exon 7 c'est à dire aux exons entourés en noir sur le sashimi plot. Le chemin du haut de la bulle, en bleu sur la figure 25b correspond à l'inclusion de l'exon 7 et le chemin du bas, en jaune sur la figure 25 correspond à l'exclusion de l'exon 7. Les arêtes sortantes du noeud 0 peuvent alors être quantifiées dans chaque condition et pour chaque réplicat. Les quantifications obtenues représenteront respectivement le nombre de lectures soutenant l'inclusion de l'exon 7 pour l'arête en bleu sur la figure 25b et la proportion de lecture soutenant l'exclusion de l'exon 7 pour l'arête en jaune sur la figure 25b. On peut alors tester si l'inclusion de l'exon 7 dans les transcrits de Pbx1 est régulée par PTBP1 (cf partie 2.1.4). On obtient une p-valeur de 0.03 et une valeur de ΔPSI égale à +0.20. Lorsque que PTBP1 est déplété, il y a 20% des lectures qui supportent l'inclusion de l'exon 7 en plus que dans la condition contrôle.

Cet événement est aussi trouvé par MAJIQ (version 2.0) [79]. Il est décrit par deux LSV (Local Splicing Variation). MAJIQ numérote les exons différemment que les auteurs du papier cité précédemment[44]. L'exon différentiellement inclus est nommé exon 8 par MAJIQ. Le premier LSV, présenté figure 26a, décrit les jonctions entre l'exon 7 et les exons 8 et 9 de Pbx1 (espérance des ΔPSI = 0.091 et -0.091). La seconde, figure 26b, décrit les jonctions entre les exons 6,7,8 et 9 (espérance des delta psi = -0.04, 0.04 et 1.5e-07). On note que MAJIQ réalise des tests concernant des jonctions qui sont présentes dans les annotations mais non présentes dans les données. Ainsi, le double saut des exons 7 et 8 est aussi testé ici. L'espérance du delta psi correspondant est, comme attendu, très faible (1.5e-07). La différence de ΔPSI obtenue entre notre méthode et MAJIQ s'explique par les différences de quantification de chacune des jonctions. MAJIQ est basées sur l'alignement des lectures sur

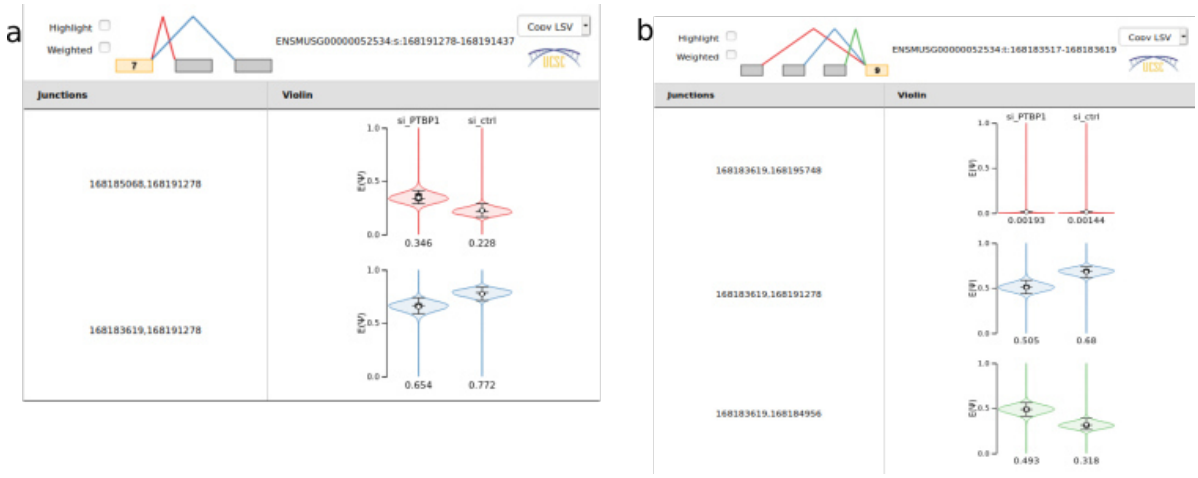


FIGURE 26 – LSVs sortis par MAJIQ ainsi que les distributions des ΔPSI correspondant au saut de l'exon 7 (nommé exon 8 par MAJIQ)

le génome de référence alors que notre méthode est basée sur l'assemblage. On quantifie ici des k-mers plutôt que les lectures supportant chacune des jonctions. Il existe des biais dans les deux cas pouvant expliquer les différences observées : possibles erreurs d'alignements ou mapping multiple dans le cas de MAJIQ, lectures chevauchant l'exon flanquant par moins de k nucléotides qui ne sont pas comptabilisées dans le cas de notre méthode.

Cet exemple est un événement d'épissage simple que l'on retrouve avec notre méthode et qui permet de confirmer des résultats connus et soutenus par une validation expérimentale mais des événements complexes sont également présents dans ce jeu de donnée.

3.1.2 Gabbr1

On s'intéresse maintenant à une autre cible connue de PTBP1 : Gabbr1. D'après [44] ce gène contient un exon dont l'inclusion est régulée par PTB. De la même façon que pour Pbx1 les lectures ont été ré-alignées sur le génome de référence de la souris (mm10) et les lectures correspondant à Gabbr1 ont été sélectionnées. La figure 27a présente le sashimi plot correspondant. Les deux premières lignes correspondent à la condition contrôle et les deux

secondes à la déplétion de PTBP1. On remarque aussi que l'intron noté i2 sur la figure peut être retenu. On observe donc la combinaison d'un saut d'exon et d'une rétention d'intron entre deux exons constitutifs (notés e1 et e3 sur la figure). Les données iclip nous informent sur les points de fixation de PTBP1 sur les transcrits de Gabbr1, ils sont annotés par des flèches rouges sur le sashimi plot.

La bulle complexe correspondante, extraite du graphe de De Bruijn est présentée figure 27c.

Le test réalisé au niveau du noeud 2 permet de mettre en évidence le rôle de PTBP1 dans la régulation de cet évènement complexe (p-valeur corrigée = 1.8e-05). le vecteur de ΔPSI associé est le suivant :

$$\Delta PSI = \begin{pmatrix} +0.32 \\ -0.32 \end{pmatrix}$$

Le saut de l'exon e2 est donc bien régulé par PTB et cela est confirmé par le test réalisé au niveau du noeud 0 (p-valeur corrigée = 2.3e-08). On retrouve donc bien ici les résultats présentés dans [44].

Notre méthode permet cependant d'identifier une bulle complexe autour de ce saut d'exon puisque l'intron i1 peut aussi être retenu. Les tests réalisés au niveau des noeuds 1 et 6 de la bulle complexe (figure 27c) permettent de tester cette rétention d'intron. Les p-valeur associées sont respectivement de 0.69 et 0.48. Cette rétention d'intron n'est donc pas régulée par PTBP1, ce résultat est discuté paragraphe 4.5.

3.2 Exemples d'évènements régulés par RED

Dans un deuxième temps, j'ai testé la méthode sur un jeu de données permettant d'étudier le rôle de RED, un facteur d'épissage connu pour être impliqué lors de l'infection des cellules humaines par le virus de la grippe. On sait que RED est recruté par le virus lors de la réplication de celui-ci et l'activité de RED est alors impactée dans les cellules infectées[17].

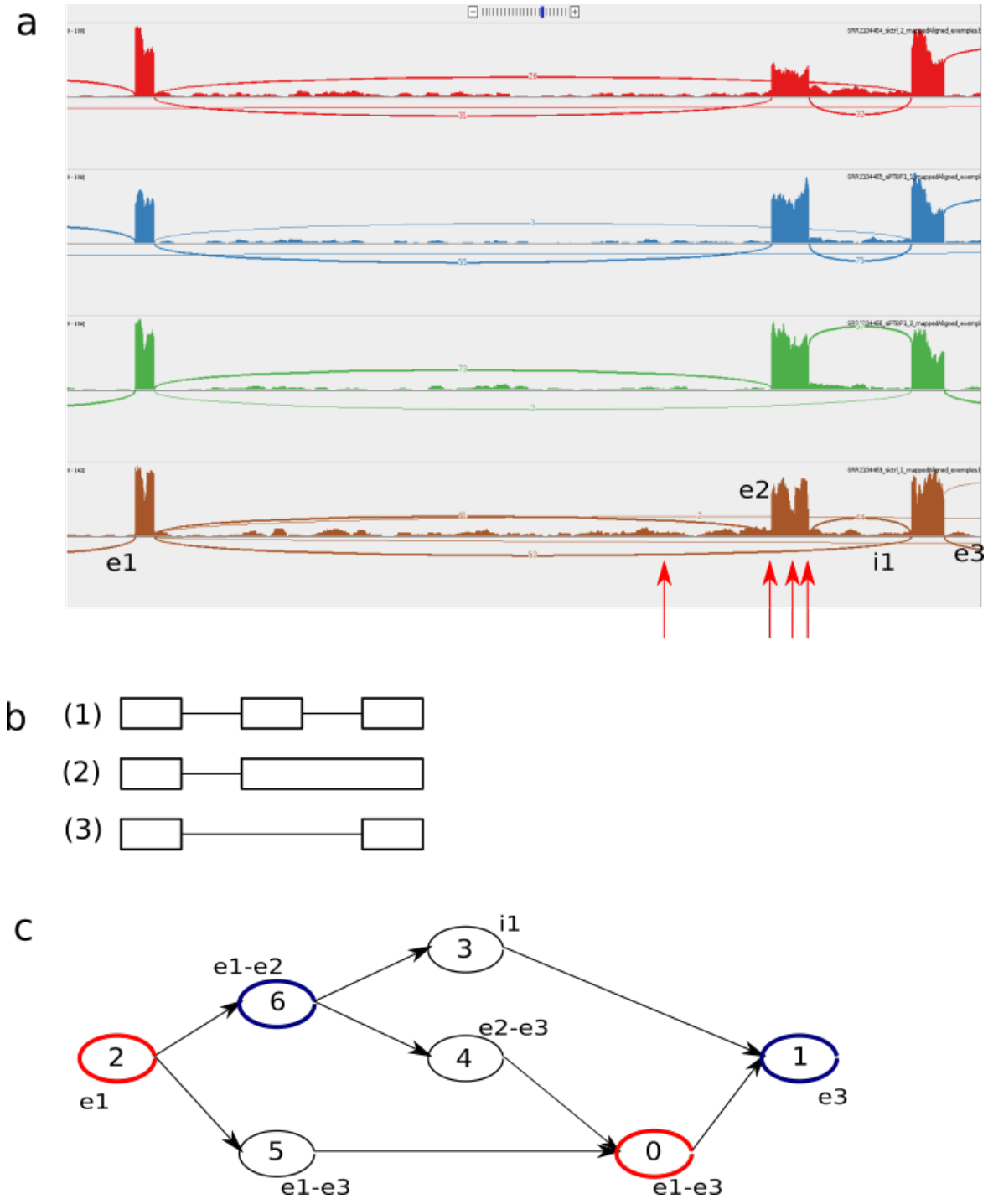


FIGURE 27 – Exemple d'une bulle complexe dans le gène Gabbr1. (a) sashimi plot, les deux premières lignes correspondent à la condition contrôle et les deux secondes à la déplétion de PTBP1. (b) Représentation des trois transcrits alternatifs. (c) Bulle complexe correspondante dans le graphe de De Bruijn

CCNT2 code pour une sous-unité du facteur d'élongation de la transcription p-TEFb. Son épissage est modifié lors de l'infection par le virus de la grippe (IAV) ainsi que lorsque RED est déplété dans des cellules infectées ou non [5]. On s'intéresse ici aux effets de la déplétion de RED dans des cellules saines.

3.2.1 CCNT2

Vue globale des variations d'épissage de CCNT2

Ce gène présente des variations d'épissage dans trois modules distincts, séparés les uns les autres par des parties constitutives. La figure 28, présente le graphe de De Bruijn obtenu en faisant tourner KisSplice sur les lectures correspondant à CCNT2 (préalablement alignées sur le génome de référence puis sélectionnées). Il existe deux étapes de filtre afin de supprimer les erreurs de séquençage avec KisSplice. La première (l'option -c), qui est un filtre absolu et permet de supprimer tous les noeuds couverts par moins de n lectures (n=2 par défaut), est réalisée avant la compression du graphe de De Bruijn. La seconde (option -C), qui permet de filter les noeuds dont la couverture relative est inférieure à 5%, à lieu après la compression du graphe. Ensuite, le graphe n'est pas compressé à nouveau. Cela explique la présence de partie linéaire, non compressées, dans le graphe de De Bruijn ainsi que celle de noeuds de longueur bien supérieure à k.

Chacun des noeuds est représenté par un camembert indiquant les quantifications dans chacune des conditions expérimentales étudiées. Ces quantifications sont obtenues en ré-alignant les lectures sur le graphe (avec BGREAT2 [43]). On compte ensuite le nombre de lectures alignées sur chacun des noeuds puis on normalise les quantifications obtenues en fonction de la taille de ceux-ci. Cette représentation permet de visualiser les variations d'épissage du gène dans leur ensemble. Depuis l'extrémité 5' des transcrits, le premier module visible correspond à un évènement d'épissage complexe, le second à un site donneur alternatif et le suivant à une rétention d'intron. Deux petites bulles sont ensuite visibles, vers l'extrémité 3' du gène, celles-ci correspondent à des SNP (single nucleotide polymorphism) dans la

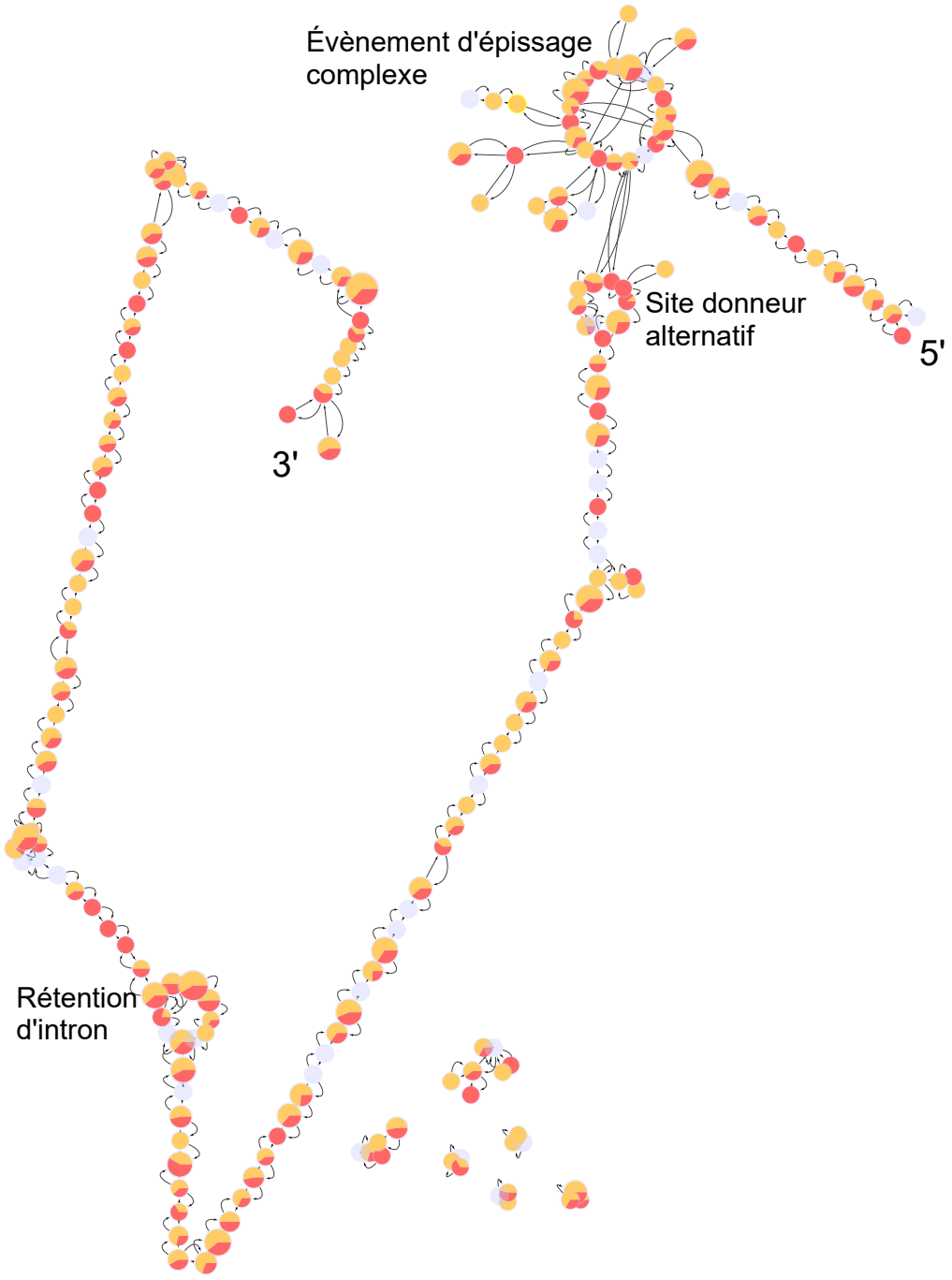


FIGURE 28 – Visualisation du graphe de De Bruijn formé par le gène CCNT2. Les variations d'épissage forment des structures dans le graphe alors que les régions constitutives à tous les transcrits correspondent aux parties linéaires du graphe. Les quantifications de chacun des noeuds sont représentées par un camembert. Les quantifications de la condition contrôle sont en jaune alors que celles de la condition où RED est déplété sont en rouge. Trois modules d'épissage sont visibles.

région du 3'UTR. Entre ces modules d'épissage, on voit des parties linéaires du graphe, elles correspondent aux régions constitutives à tous les transcrits.

Dans la suite on se focalise sur l'évènement d'épissage complexe cité ci-dessus et à la rétention d'intron proche de l'extrémité 3' des transcrits.

Épissage complexe

Ce gène présente des variations d'épissage complexe, le sashimi plot correspondant est présenté figure 29a. Les exons notés e1 et e4 sont constitutifs à tous les transcrits de CCNT2. L'exon e2 peut être exclu des transcrits et l'exon e3 existe dans une version courte (notée e3') et une version longue (e3). Cet exon peut aussi être inclus ou exclu des transcrits étudiés.

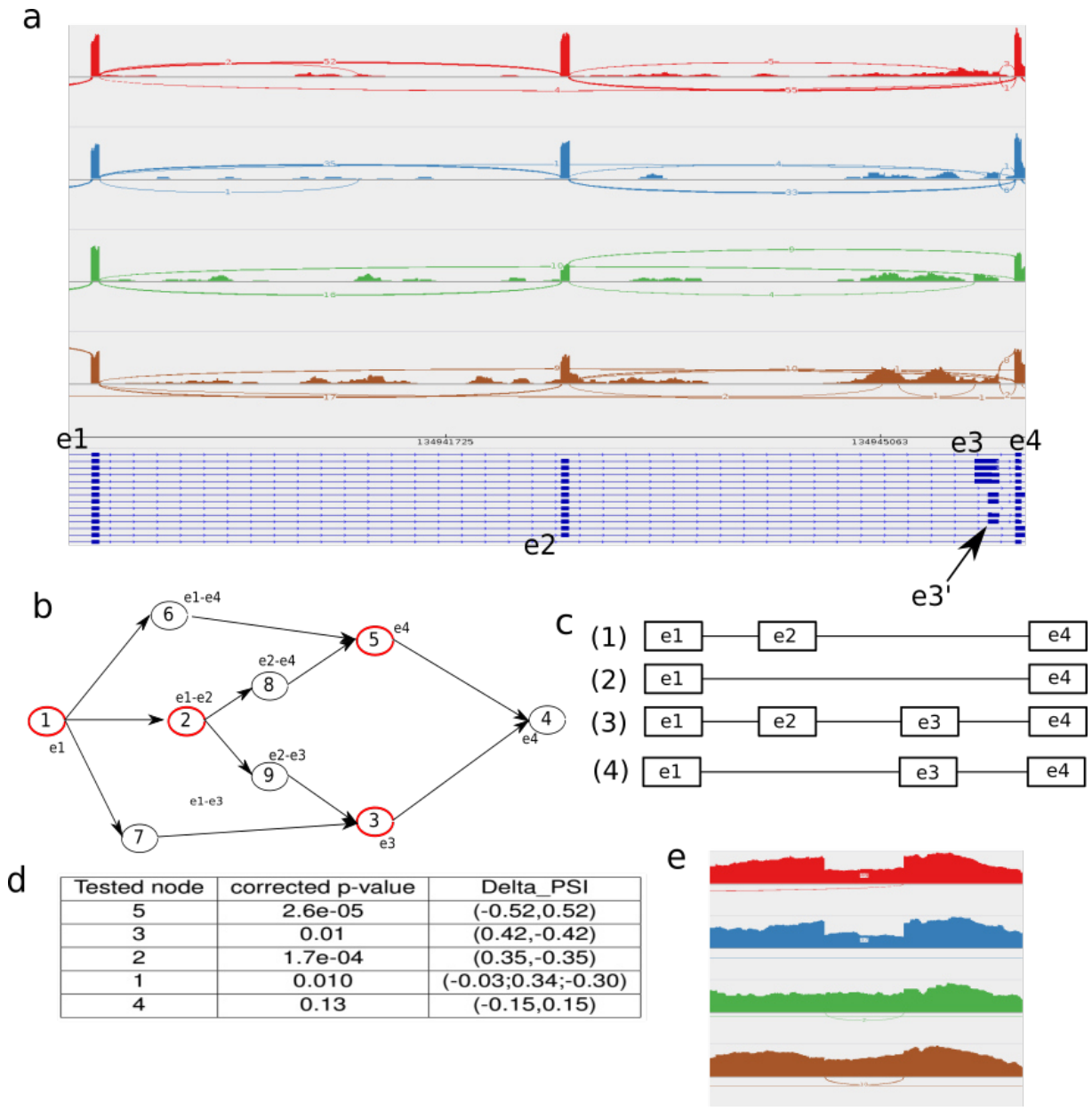
La bulle complexe correspondante est présentée Figure 29b. Les noeuds 1 et 4 correspondent aux exons constitutifs c'est-à-dire respectivement aux exons e1 et e4. Les noeuds testés sont entourés en noir et ceux dont la p-valeur corrigée est inférieure à 0.05 sont entourés en rouge. Les valeurs des p-valeurs ainsi que des ΔPSI obtenus sont indiquées figure 29d.

Le noeud 1 (figure 29b) correspond à l'exon e1. Une fois les arêtes à tester sélectionnées (voir paragraphe 2.3), il y a trois possibilités d'appariement, à partir du site d'épissage donneur de l'exon e1, avec un site d'épissage accepteur possible :

1. Avec le site accepteur de l'exon e3, (exclusion de l'exon e2 mais inclusion de l'exon e3, transcrit (4) figure29c)
2. Avec le site accepteur de l'exon e2 (inclusion de l'exon e2, transcrit (2) ou (4) figure29c)
3. Avec le site accepteur de l'exon e4 (double saut des exons e2 et e3, transcrit (2)figure29c)

La p-valeur correspondante est de 0.01 et les ΔPSI associées sont les suivants :

$$\Delta PSI = \begin{pmatrix} -0.03 \\ +0.34 \\ -0.30 \end{pmatrix}$$



Tested node	corrected p-value	Delta_PSI
5	2.6e-05	(-0.52,0.52)
3	0.01	(0.42,-0.42)
2	1.7e-04	(0.35,-0.35)
1	0.010	(-0.03;0.34;-0.30)
4	0.13	(-0.15,0.15)

FIGURE 29 – Exemple d'évènement complexe pour le gène CCNT2. CCNT2 contient un événement complexe aux positions génomiques suivantes : chr2 : 134,938,978 - 134,946,156 a) Sashimi plot, les deux première lignes correspondent à la condition contrôle et les deux dernières à la déplétion de RED (b) Bulle complexe correspondante dans le graphe de De Bruijn (c) Représentation des transcrits alternatifs obtenus. (d) P-valeurs corrigées et ΔPSI obtenus pour chacun des noeuds testés. (e) Sashimi plot correspondant au petit intron régulé par RED , il correspond à un deuxième évènement d'épissage.

Les variations d'épissage observées ici sont donc bien régulées par RED. Pour comprendre plus précisément ce qu'il se passe, on peut s'intéresser au ΔPSI et donc à la magnitude de l'effet observé pour chaque possibilité d'appariement citée ci-dessus. La valeur absolue de ΔPSI la plus faible (-0.03) correspond à la proposition 1 ci-dessus c'est-à-dire à l'inclusion de l'exon e3. Cela est cohérent avec le sashimi plot (figure 29a) où l'on peut voir que e3 est très peu couvert quelle que soit la condition expérimentale. Les ΔPSI plus élevés (+0.34 et -0.30) correspondent respectivement aux propositions 2 et 3 ci-dessus. L'effet est donc bien plus marqué pour le saut de l'exon e2 que pour celui de l'exon e3.

À l'intérieur de la bulle complexe on réalise d'autres tests locaux et trois d'entre eux permettent d'identifier des régions génomiques différentiellement épissées (Figure 29b). Au niveau du noeud 2, on teste le saut de l'exon e3 dans le cas où l'exon e2 est inclus c'est à dire avec les exons e2 et e4 comme exons flanquants. Concernant le noeud 5, les deux arêtes entrantes considérées correspondent au saut de l'exon e2 si l'exon e3 n'est pas inclus. Enfin, le test réalisé pour le noeud 3 correspond au saut de l'exon e2, considérant les exons dans le cas ou l'exon e3 est retenu. La p-valeur est plus importante que celle du noeud 5 alors que l'on teste dans les deux cas l'inclusion de l'exon e2 ($0.01 > 1.7e-04$) mais cela s'explique par la faible couverture de l'exon e3. Bien que l'on considère ici l'évènement complexe dans sa globalité, en testant localement tous les noeuds branchants de la bulle complexe on teste plusieurs fois l'inclusion des exons e2 et e3 et cela maintient, en partie, la redondance observée dans les résultats de KisSplice. Ce point est discuté paragraphe 4.3.

Rétention d'un petit intron

Plus loin vers l'extrémité 5' du gène CCNT2, on observe aussi une rétention d'intron. C'est un événement d'épissage simple mais de la même façon que pour le gène Pbx1 nous pouvons faire une hypothèse sur les mécanismes de régulation en jeu. En effet, RED est connu pour aider à la reconnaissance des petits introns [30]. L'intron présenté ici a une longueur de 102

paires de bases et on peut donc supposer que sa rétention est, au moins en partie, régulée par RED. Le sashimi plot correspondant à cet événement d'épissage est représenté figure 29e. Les deux premières lignes correspondent à la condition contrôle et les deux secondes à la déplétion de RED. Quand RED n'est pas là l'intron est bien plus retenu, l'intron est donc mieux reconnu en condition contrôle. Ces observations sont confirmées par les résultats obtenus avec la méthode présentée ici. On retrouve bien une bulle simple correspondant à cette rétention d'intron. La p-valeur associée est égale à 0.154e10-4 et le ΔPSI de 0.33. Ainsi, nos résultats confirment donc l'hypothèse que RED aide à la reconnaissance de cet intron.

Comparaison avec les résultats de MAJIQ

MAJIQ [79] trouve 4 LSV (Local Splicing Variation) qui correspondent tous à des évènements d'épissage simple. L'un d'entre eux correspond à la rétention du petit intron présenté ci-dessus (espérance du ΔPSI égale à 0.29). Ensuite, le deuxième LSV correspond au saut de l'exon e2 (espérance du ΔPSI égale à 0.21) et le troisième LSV correspond à un site donneur alternatif, présent juste après la bulle complexe présentée ci dessus (voir paragraphe 4.1 de la discussion). L'espérance du ΔPSI associé à cet événement est égale à 0.22. Le saut de l'exon e3 n'est pas testé, les filtres sur les comptages faibles appliqués par MAJIQ étant plus stringents que les nôtres.

Enfin, MAJIQ teste aussi un évènement dont les jonctions sont annotées mais non présentes dans les données. Sans génome de référence, on ne peut assembler que ce qui est présent dans les données et donc uniquement ce qui est transcrit. Cependant, les transcriptomes étant tissu-spécifiques, il semble ici pertinent de ne tester que ce qui est observé dans les données plutôt que tout ce qui a déjà été vu, quelque soit le tissu considéré. Prendre en compte les jonctions annotées mais non couvertes fait augmenter le nombre de tests réalisés et donc diminuer la puissance statistique. Notre méthode ne prend donc pas en compte les jonctions qui sont absentes des données étudiées mais présentes dans les annotations.

3.2.2 CLK1

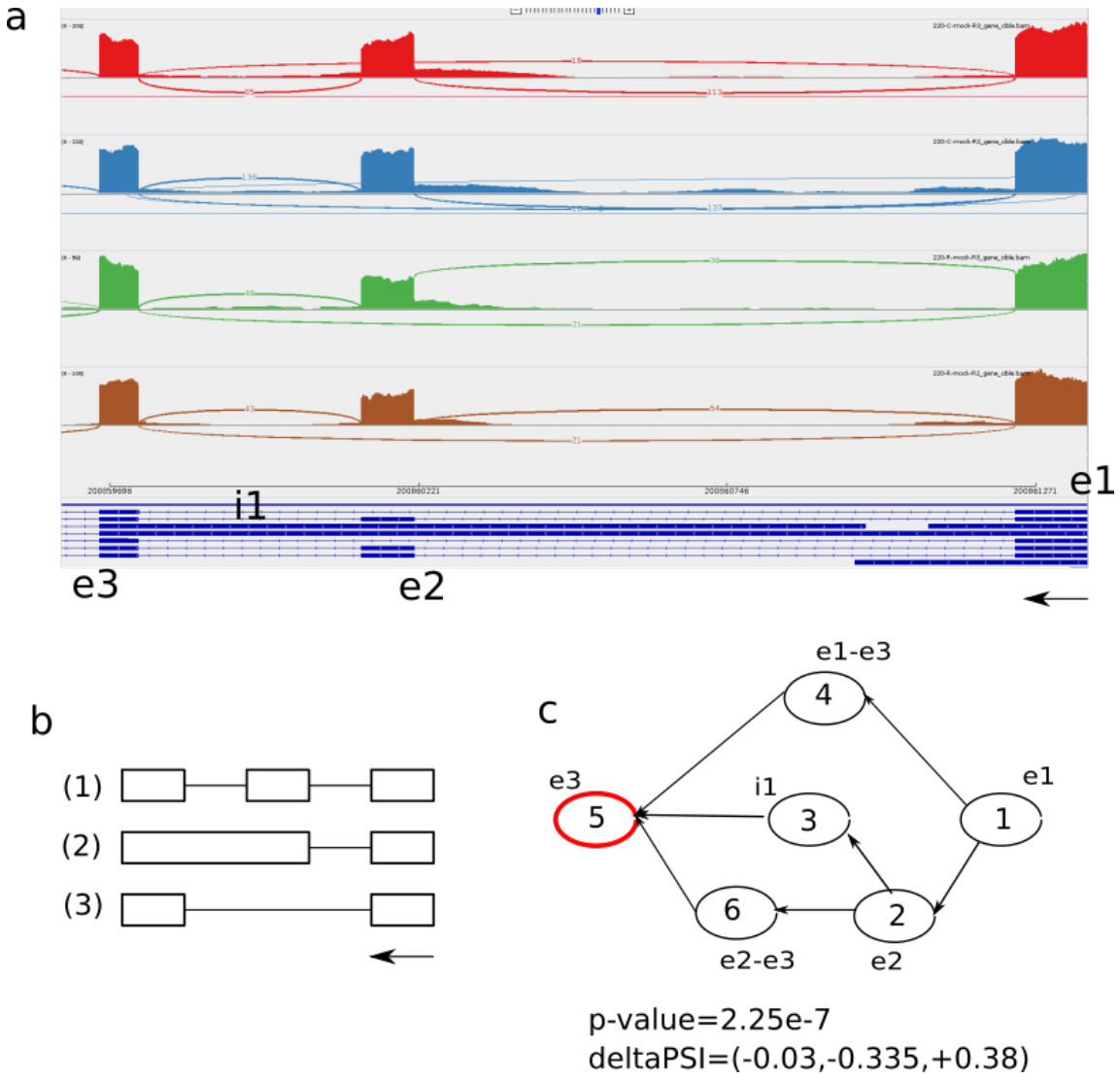


FIGURE 30 – Exemple du gène CLK1 qui contient un évènement complexe aux positions génomiques suivantes : chr2 : 200,859,679 - 200,861,283 (a) sashimi plot, les deux premières lignes correspondent à la condition contrôle et les deux dernières à la déplétion de RED. (b) Représentation des trois transcrits obtenus (c) Bulle complexe correspondante dans le graphe de De Bruijn.

Le gène CLK1 contient un événement complexe composé d'un saut d'exon et d'une rétention d'intron. Le sashimi plot correspondant est présenté figure 30a. Pour cette région

génomique, KisSplice considère 3 évènements pairwise. On considère ici 3 transcrits différents, représentés sur la figure 30b. Dans la condition contrôle, le transcrit majoritaire est celui incluant l'exon e2 c'est-à-dire le transcrit (1) sur la figure 30. Avec le modèle classique de KisSplice on compare les transcrits deux à deux. Il est intéressant de noter que ces comparaisons peuvent induire, en plus d'une redondance importante dans les transcrits considérés, des erreurs d'analyse en aval et c'est le cas ici. Un des post-traitements couramment réalisé sur les sorties de KisSplice, lorsque cela est possible, est de ré-aligner les bulles sur un génome de référence puis d'analyser ces alignements afin d'en déduire notamment le sous-type d'événements d'épissage alternatif (saut d'exon, rétention d'intron, site donneur ou accepteur alternatif). Ces analyses ont été automatisées dans un paquet nommé KisSplice2refGenome (K2RG). Pour conclure sur le type d'évènement d'épissage, K2RG considère le nombre de blocs alignés pour chacun des chemins de la bulle. Par exemple, dans le cas ou le chemin du haut ainsi que le chemin du bas s'alignent chacun en deux blocs sur le génome de référence mais que le deuxième bloc du chemin du haut est plus long que celui du chemin du bas alors on conclut à un site d'épissage accepteur alternatif (voir le manuel de KissPlice2refGenome, page 10, pour plus de détails. http ://kissplice.prabi.fr/tools/kiss2refgenome).

Ici, lorsque l'on considère la comparaison des transcrits (2) et (3) (figure 30) entre eux, c'est-à-dire le chemin incluant l'intron ainsi que le chemin contenant uniquement la jonction entre les deux exons flaquants (c'est-à-dire correspondant au saut d'exon), on rentre dans le cas décrit ci-dessus (figure 30b). KisSplice2refGenome considère donc cet évènement comme un site d'épissage accepteur alternatif alors que, le transcrit (1) étant abondant quelle que soit la condition, il n'en est rien.

Ainsi, comparer ces deux isoformes, indépendamment des autres, n'est pas cohérent d'un point de vue biologique. L'exon présent dans le transcrit (1) doit être considéré, d'autant plus qu'ici le transcrit (1) est majoritaire en condition contrôle. Considérant la présence de cet exon, le transcrit (2) correspond bien à une rétention d'intron et il est pertinent de l'étudier comme tel. Le modèle pairwise est donc ici clairement insuffisant pour décrire cet

événement d'épissage et considérer les variations d'épissage observées ici comme un seul événement complexe semble bien plus pertinent d'un point de vue biologique.

La figure 30c présente la bulle complexe correspondant à cet événement. Le noeud 5 correspond à l'extrémité 5' de l'exon annoté comme e3 sur la figure 30a. Lors de l'analyse différentielle, le test réalisé au niveau de ce noeud permet de considérer les trois transcrits décrits plus haut. Il y a en effet 2 possibilités de site d'épissage donneur à apparier avec le site accepteur de l'exon e1 :

— avec le site donneur de l'exon e1 (saut de l'exon e2)

— avec le site donneur de l'exon e2 (e2 inclus)

De plus, le site accepteur de l'exon e3, ainsi que le site donneur de l'exon e2, peuvent ne pas être considérés par le spliceosome et l'intron sera alors retenu.

La p-valeur corrigée obtenue lors de ce test est de 2.25e-07 et les ΔPSI correspondant sont les suivants : $\Delta PSI = \begin{pmatrix} -0.033 \\ -0.353 \\ +0.387 \end{pmatrix}$

D'après notre méthode, les variations d'épissage observées sont bien régulées par RED. Les valeurs de ΔPSI représentent la magnitude de l'effet pour chacune des possibilités étudiées. La première valeur, plus faible que les autres (-0.033) correspond à la rétention d'intron. L'effet ici est donc très faible, et c'est plutôt le saut de l'exon e2 qui semble justifier la faible p-valeur observée. Ces conclusions sont confirmées par le sashimi plot (figure 30a), l'exon e2 est moins couvert lorsque RED est déplété et RED semble donc aider à sa reconnaissance.

Cet événement complexe correspond à deux LSV sortis par MAJIQ. Le premier permet de tester le saut de l'exon e2 ainsi que le double saut des exons e3 et e2 qui est présent dans les annotations mais que nous ne retrouvons pas dans les données et qui n'est donc pas vu sans génome de référence. Les espérances des ΔPSI associés sont les suivantes : $\mathrm{E}(\Delta PSI) = \begin{pmatrix} 0 \\ 0.255 \\ -0.258 \end{pmatrix}$

Le second LSV correspond à la rétention de l'intron i1. L'espérance du ΔPSI est égale à 0.072.

Concernant cet évènement complexe, on retrouve donc bien des résultats équivalents aussi bien avec MAJIQ que sans génome de référence.

4 | Discussion et perspectives

On propose ici une méthode d'analyse des événements d'épissage complexe sans génome de référence ni annotation. Celle-ci est basée sur l'identification de bulles complexes dans le graphe de De Bruijn induit par les bulles simples sorties par KisSplice. KisSplice permet généralement d'identifier plusieurs centaines et parfois plusieurs milliers, selon les jeux de données, d'événements d'épissage simple dont environ un tiers peuvent êtres regroupés au sein d'évènements d'épissage complexe. Ainsi, dans les cas où l'on observe plus de deux transcrits possibles entre deux sites d'épissage constitutifs, on évite de comparer systématiquement les transcrits deux à deux et l'on supprime une partie de la redondance observée dans les sorties de KisSplice. Cela devrait faciliter le travail des biologistes pour l'analyse approfondie des gènes d'intérêts pour lesquels le modèle actuel de KisSplice n'est pas suffisant.

4.1 Échelle d'étude intermédiaire pour étudier l'épissage

En observant plusieurs dizaines d'exemples, on se rend compte que souvent, les différents transcrits d'un même gène sont composés de parties constitutives, c'est-à-dire communes à tous les transcrits, mais l'on observe aussi des variations d'épissage entre ces parties constitutives. Ainsi pour un gène, on observe des zones où les sites d'épissage sont plus forts et d'autre où ils sont plus faibles et où les variations sont plus nombreuses. Pour définir ces régions plus variables, que l'on décide de nommer ici module d'épissage, on peut les délimiter par les sites d'épissage constitutifs qui les entourent. Ces modules sont visibles sur la figure 28 concernant le gène CCNT2. On observe trois module d'épissage : un événements

complexe, un site donneur alternatif et une rétention d'intron. On propose ici une échelle d'étude intermédiaire entre les échelles locale (centrée sur l'exon) et globale (reconstruction de transcrit complet) pour l'étude de l'épissage. On choisit de considérer comme un module d'épissage toutes les variations comprises entre deux sites constitutifs à tous les transcrits d'un gène. On construit ainsi des événements d'épissage complexe correspondant à chacun de ses modules. En effet, reconstruire les transcrits complets depuis les lectures courtes est un problème difficile et tandis que les lectures longues ne sont pas toujours complètes (voir chapitre 1) les méthodes actuelles basées sur les lectures courtes font encore beaucoup d'erreurs [71]. D'un point de vue mécanistique, l'épissage est co-transcriptionnel et a lieu de l'extrémité 5' à l'extrémité 3' du transcrit [9]. On s'attend alors à ce que les exons soient reconnus au fur et à mesure que la transcription a lieu. Ainsi il parait cohérent de considérer les modules d'épissage indépendamment les uns des autres.

Cependant, lorsque deux évènements d'épissage complexe partagent un exon constitutif (la fin d'un événement correspond au site accepteur et le début de second au site donneur de l'exon) comme c'est le cas pour le gène CCNT2 (figure 29) il pourrait être intéressant de relier entre eux les deux événements.

Par exemple, la fin de la bulle complexe du gène CCNT2 présentée plus haut correspond aussi au début de la bulle simple suivante (figure 28). Si l'on suppose que RED aide à la reconnaissance de l'exon e4 alors il serait plus cohérent de considérer le site donneur alternatif de l'exon e5 comme faisant partie de l'événement d'épissage complexe cité plus haut. Cette proposition n'est pas encore implémentée mais c'est une perspective qui me semble intéressante pour améliorer notre modèle.

4.2 Puissance statistique

On considère ici les quantifications des arêtes du graphe de De Bruijn plutôt que celles des noeuds dans le modèle classique de KisSplice. Ces quantifications correspondent donc au nombre de fois ou le (k+1)-mer correspondant à l'arête quantifiée est présent dans les

lectures. Ainsi on ne se sert pas des lectures exoniques pour réaliser nos tests. On perd donc en puissance statistique par rapport à la version classique de KisSplice. Cependant, quantifier les arêtes nous permet de tester de façon très locale -au niveau du site d'épissage- les choix réalisés. On peut ainsi tester, pour un site d'épissage, l'appariement avec le deuxième site d'épissage de l'intron considéré et donc l'usage des introns. On réalise donc plusieurs tests locaux au sein d'un même événement d'épissage complexe afin d'identifier les régions génomiques différentiellement incluses.

Cette méthode est moins puissante d'un point de vue statistique pour les événements simples pour lesquels le modèle actuel de KisSplice est adapté. Ainsi l'analyse proposée est un post-traitement de KisSplice afin de détecter les cas où le modèle actuel n'est pas suffisant pour décrire les variations d'épissage observées.

4.3 Comparaison de transcrits mineurs entre eux

Comme mentionné lors de l'étude du gène CCNT2 (paragraphe 3.2.1), certains tests réalisés à l'intérieur des bulles complexes sont, tout de même, redondants. En effet, pour certains événements complexes, on teste plusieurs fois l'inclusion d'une même région génomique. Cela vient du fait que l'on réalise un test pour chaque noeud de degré sortant ou entrant supérieur ou égal à 2. Cela implique aussi que l'on teste parfois l'inclusion d'une région génomique en faisant varier les exons flanquants et peut conduire à la comparaison de transcrits mineurs entre eux. Ce problème est aussi observé lors des comparaisons pairwise effectuées avec KisSplice ainsi qu'avec MAJIQ puisque tous les LSV (local splicing variations) -non filtrés pour les comptages faibles- font l'objet d'un test.

Une possibilité simple pour réduire la redondance et le nombre de tests réalisés serait de ne tester que les noeuds de degré sortant supérieur ou égal à 2. Cela permettrait de réaliser un seul test au lieu de deux pour une bulle simple -imbriquée dans une bulle complexe plus grande ou non- par exemple. Cependant il existe de nombreux cas que l'on ne traiterait pas. Les sites donneurs alternatifs par exemple seraient systématiquement ignorés. Aussi, cela

poserait problème pour certains cas d'événements complexes observés régulièrement comme ceux composés d'un saut d'exon et d'une rétention d'intron.

4.4 Questions mécanistiques

On sait que le splicoesome peut s'assembler autour des introns mais aussi d'abord autour des exons dans certains cas (paragraphe 2.2.2 de l'introduction). Notre méthode permet de construire des hypothèses concernant ce qui est reconnu lors de la régulation de l'épissage. On peut par exemple, pour le gène CCNT2, supposer que l'exon e2 (Figure 29) dont l'inclusion est régulé par RED est mieux reconnu lorsque RED est présent. Il semble aussi cohérent de supposer que le spliceosome s'assemble autour du petit intron qui est retenu, un peu plus loin le long de ce même gène, et dont l'inclusion est aussi régulée par RED. Cette supposition est d'autant plus cohérente que l'on sait que RED aide à reconnaître les petits introns.

Tout cela reste, bien entendu, des suppositions, seules des validations expérimentales permettraient de les confirmer mais notre méthode permet de poser de telles hypothèses et pourrait ainsi aider les biologistes à débattre de ces questions encore ouvertes sur des exemples précis.

4.5 Régulation de l'épissage et bruit de la machinerie d'épissage

Toutes les variations d'épissage ne sont pas régulées dans la cellule (voire paragraphe 2.4.1 de l'introduction) [77]. Le fait qu'un évènement d'épissage soit régulé entre deux conditions expérimentales est un argument pour affirmer que les variations observées ne sont pas dues au hasard, d'autant plus lorsque l'on étudie la déplétion d'un facteur d'épissage. Notre méthode permet de mettre en évidence de tels exemples mais il existe aussi des cas où l'on ne peut pas prouver que l'épissage est régulé. C'est le cas de la rétention d'intron du gène Gabbr1 (voir paragraphe 3.1.2).

Pour Gabbr1, la bulle simple correspondant au saut de l'exon est régulé par PTBP1 mais ce n'est pas le cas de la rétention d'intron. Ainsi on sort un évènement complexe, plutôt qu'un évènement simple, uniquement dû au bruit de la machinerie d'épissage. Dans ce cas précis, le modèle pairwise est adapté à l'étude du saut de l'exon e2 mais notre méthode permet aussi de mettre en évidence des arguments permettant de discuter de la question, encore ouverte, de la prévalence des évènements d'épissage régulés.

Ainsi, la complexité importante de certains évènements peut être débattue si une grande partie des variations d'épissage observées sont dus au bruit de la machinerie d'épissage. Plus la profondeur de séquençage est importante, plus on a accès aux transcrits rares et plus les variations d'épissage dues au hasard seront visibles. Ainsi plus on augmente la profondeur de séquençage et plus les évènements observés seront complexes. Par exemple, pour Gabbr1, si la profondeur choisie n'avait pas permis d'observer la rétention de l'intron i1, le modèle de comparaison deux à deux des transcrits aurait été suffisant. Le choix du seuil à fixer pour filtrer les transcrits rares n'est pas simple à trancher. Notre méthode permet, ici aussi, de mettre en évidence des arguments pour discuter cette question.

4.6 Perspectives concernant les lectures longues

La longueur des lectures issues des technologies de séquençage de troisième génération permet de résoudre les questions de couplage des exons et plus largement des évènements d'épissage complexes entre eux.

Une perspective intéressante de ce travail serait d'intégrer les lectures longues à l'analyse effectuée. En effet, chacun des transcrits correspond à un chemin dans le graphe de De Bruijn mais tous les chemins ne correspondent pas à des transcrits effectivement vus dans les données. En alignant les lectures longues sur les sous graphes des bulles complexes on pourrait identifier les chemins effectivement vus dans les données, au moins pour les transcrits les plus fréquents. Cependant la profondeur des technologies de troisième génération augmentant

rapidement ces derniers mois, de plus en plus de transcrits rares sont vus avec les lectures longues et on peut s'attendre à avoir, dans un futur proche, accès à tous les transcrits vu en illumina dans les données des lectures longues.

À une échelle plus large, on pourrait aussi réaligner les lectures longues directement sur le graphe induit par les bulles simples de KisSplice afin de phaser les évènements complexes entre eux et d'identifier le ou les transcrits majoritaires à l'échelle du gène. Cependant, il ne me semble pas pertinent d'essayer de phaser les évènements entre eux si aucun des événements détectés pour le gène n'est régulé. En effet, si, pour un gène particulier, les différents transcrits observés résultent du bruit de la machinerie d'épissage alors on ne s'attend pas à ce que l'inclusion des différents exons entre eux soient couplés.

Bibliographie

[1] Mohamad Al kadi, Nicolas Jung, Shingo Ito, Shoichiro Kameoka, Takashi Hishida, Daisuke Motooka, Shota Nakamura, Tetsuya Iida, and Daisuke Okuzaki. UNAGI : an automated pipeline for nanopore full-length cDNA sequencing uncovers novel transcripts and isoforms in yeast. *Functional and Integrative Genomics*, 20(4) :523–536, 7 2020.

[2] Israa Alqassem, Yash Sonthalia, Erika Klitzke-Feser, Heejung Shim, and Stefan Canzar. McSplicer : a probabilistic model for estimating splice site usage from RNA-seq data. *Bioinformatics*, 1 2021.

[3] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome research*, 22(10) :2008–17, 10 2012.

[4] Usama Ashraf, Clara Benoit-Pilven, Vincent Lacroix, Vincent Navratil, and Nadia Naffakh. Advances in Analyzing Virus-Induced Alterations of Host Cell Splicing, 3 2019.

[5] Usama Ashraf, Clara Benoit-Pilven, Vincent Navratil, Cécile Ligneau, Guillaume Fournier, Sandie Munier, Odile Sismeiro, Jean-Yves Coppée, Vincent Lacroix, and Nadia Naffakh. Influenza virus infection induces widespread alterations of host cell splicing. *NAR Genomics and Bioinformatics*, 2(4), 11 2020.

[6] Ergude Bao and Lingxiao Lan. HALC : High throughput algorithm for long read error correction. *BMC Bioinformatics*, 18(1), 4 2017.

[7] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1) :289–300, 1 1995.

[8] Clara Benoit-Pilven, Camille Marchet, Emilie Chautard, Leandro Lima, Marie-Pierre Lambert, Gustavo Sacomoto, Amandine Rey, Audric Cologne, Sophie Terrone, Louis Dulaurier, Jean-Baptiste Claude, Cyril F. Bourgeois, Didier Auboeuf, and Vincent Lacroix. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Scientific Reports*, 8(1) :4307, 12 2018.

[9] David Bentley. The mRNA assembly line : Transcription and processing machines in the same factory, 6 2002.

[10] Susan M. Berget. Exon recognition in vertebrate splicing, 2 1995.

[11] Elsa Bernard, Laurent Jacob, Julien Mairal, and Jean Philippe Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, 30(17) :2447–2455, 2014.

[12] Grzegorz M. Boratyn, Jean Thierry-Mieg, Danielle Thierry-Mieg, Ben Busby, and Thomas L. Madden. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*, 20(1) :405, 7 2019.

[13] Elena Bushmanova, Dmitry Antipov, Alla Lapidus, and Andrey D. Prjibelski. RnaSPAdes : A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8(9) :1–13, 9 2019.

[14] Ivan de la Rubia, Joel Indi, Silvia Carbonell-Sala, Julien Lagarde, M Mar Albà, and Eduardo Eyras. Reference-free reconstruction and quantification of transcriptomes from Nanopore long-read sequencing. *bioRxiv*, page 2020.02.08.939942, 5 2020.

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

[16] Naima Ahmed Fahmi, Heba Nassereddeen, Jae Woong Chang, Meeyeon Park, Hsin Sung Yeh, Jiao Sun, Deliang Fan, Jeongsik Yong, and Wei Zhang. As-quant : Detection and visualization of alternative splicing events with rna-seq data. *International Journal of Molecular Sciences*, 22(9), 5 2021.

[17] Guillaume Fournier, Chiayn Chiang, Sandie Munier, Andru Tomoiu, Caroline Demeret, Pierre-Olivier Vidalain, Yves Jacob, and Nadia Naffakh. Recruitment of RED-SMU1 Complex by Influenza A Virus RNA Polymerase to Control Viral mRNA Splicing. *PLoS Pathogens*, 10(6) :e1004164, 6 2014.

[18] Daniel R. Garalde, Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E. Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J. Heron, and Daniel J. Turner. Highly parallel direct RN A sequencing on an array of nanopores. *Nature Methods*, 15(3) :201–206, 3 2018.

[19] Manuel Garber, Manfred G. Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq, 6 2011.

[20] Manuel L. Gonzalez-Garay. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). pages 141–160. Springer, Dordrecht, 2016.

[21] Mar Gonzàlez-Porta, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7) :R70, 7 2013.

[22] Manfred G. ; Grabherr, Dawn A. Thompson Ido Amit Xian Adiconis Lin Fan Raktima Raychowdhury Qiandong Zeng Zehua Chen Evan Mauceli Nir Hacohen Andreas Gnirke Nicholas Rhind Federica di Palma Bruce W. Nir Brian J. Haas, Moran Yassour Joshua Z. Levin, Friedman, and Aviv Regev. Trinity : reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7) :644–652, 2013.

[23] Mitchell Guttman, Manuel Garber, Joshua Z. Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J. Koziol, Andreas Gnirke, Chad Nusbaum, John L. Rinn, Eric S. Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5) :503–510, 5 2010.

[24] Anna Hegele, Atanas Kamburov, Arndt Grossmann, Chrysovalantis Sourlis, Sylvia Wowro, Mareike Weimann, Cindy L. Will, Vlad Pena, Reinhard Lührmann, and Ulrich Stelzl. Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome. *Molecular Cell*, 45(4) :567–580, 2 2012.

[25] Lydia Herzel, Diana S.M. Ottoz, Tara Alpert, and Karla M. Neugebauer. Splicing and transcription touch base : Co-transcriptional spliceosome assembly and function, 10 2017.

[26] Yu Hu, Li Fang, Xuelian Chen, Jiang F Zhong, Mingyao Li, and Kai Wang. LIQA : Long-read Isoform Quantification and Analysis. *bioRxiv*, page 2020.09.09.289793, 4 2021.

[27] José María Izquierdo, Nuria Majós, Sophie Bonnal, Concepción Martínez, Robert Castelo, Roderic Guigó, Daniel Bilbao, and Juan Valcárcel. Regulation of fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Molecular Cell*, 19(4) :475–484, 8 2005.

[28] Henrik Kaessmann, Nicolas Vinckenbosch, and Manyuan Long. RNA-based gene dupli-

cation : mechanistic and evolutionary insights. *Nature reviews. Genetics*, 10(1) :19–31, 1 2009.

[29] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12) :1009–15, 12 2010.

[30] Sandra Keiper, Panagiotis Papasaikas, Cindy L. Will, Juan Valcárcel, Cyrille Girard, and Reinhard Lührmann. Smu1 and RED are required for activation of spliceosomal B complexes assembled on short introns. *Nature Communications*, 10(1) :1–15, 12 2019.

[31] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution : diversification, exon definition and function. *Nature Reviews Genetics*, 11(5) :345–355, 5 2010.

[32] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT : A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4) :357–360, 3 2015.

[33] Eddo Kim, Alon Magen, and Gil Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1) :125–131, 1 2007.

[34] Sergey Koren, Michael C. Schatz, Brian P. Walenz, Jeffrey Martin, Jason T. Howard, Ganeshkumar Ganapathy, Zhong Wang, David A. Rasko, W. Richard McCombie, Erich D. Jarvis, and Adam M. Phillippy. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7) :693–700, 7 2012.

[35] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu : Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, 27(5) :722–736, 5 2017.

[36] Sam Kovaka, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg,

and Mihaela Pertea. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20(1) :278, 12 2019.

[37] Liana F. Lareau, Angela N. Brooks, David A.W. Soergel, Q. Meng, and Steven E. Brenner. The coupling of alternative splicing and nonsense-mediated mRNA decay., 2007.

[38] Hervé Le Hir, David Gatfield, Elisa Izaurralde, and Melissa J. Moore. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO Journal*, 20(17) :4987–4997, 9 2001.

[39] Heng Li. Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18) :3094–3100, 9 2018.

[40] Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1) :151–158, 1 2018.

[41] Leandro Lima, Camille Marchet, Ségolène Caboche, Corinne da Silva, Benjamin Istace, Jean Marc Aury, Hélène Touzet, and Rayan Chikhi. Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data, 7 2019.

[42] Leandro Lima, Blerina Sinaimeri, Gustavo Sacomoto, Helene Lopez-Maestre, Camille Marchet, Vincent Miele, Marie France Sagot, and Vincent Lacroix. Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorithms for Molecular Biology*, 12(1), 2 2017.

[43] Antoine Limasset, Bastien Cazaux, Eric Rivals, and Pierre Peterlongo. Read mapping on de Bruijn graphs. *BMC bioinformatics*, 17(1) :237, 2016.

[44] Anthony J. Linares, Chia Ho Lin, Andrey Damianov, Katrina L. Adams, Bennett G. Novitch, and Douglas L. Black. The splicing regulator PTBP1 controls the activity of the

transcription factor Pbx1 during neuronal differentiation. *eLife*, 4(DECEMBER2015), 12 2015.

[45] Hélène Lopez-Maestre, Lilia Brinza, Camille Marchet, Janice Kielbassa, Sylvère Bastien, Mathilde Boutigny, David Monnin, Adil El Filali, Claudia Marcia Carareto, Cristina Vieira, Franck Picard, Natacha Kremer, Fabrice Vavre, Marie France Sagot, and Vincent Lacroix. SNP calling from RNA-seq data without a reference genome : Identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research*, 44(19), 11 2016.

[46] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12) :550, 12 2014.

[47] Mohammed Amin Madoui, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean Marc Aury. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics*, 16(1), 12 2015.

[48] Lynne E. Maquat. Nonsense-mediated mRNA decay : Splicing, translation and mRNP dynamics, 2 2004.

[49] Camille Marchet, Lolita Lecompte, Corinne Da Silva, Corinne Cruaud, Jean Marc Aury, Jacques Nicolas, and Pierre Peterlongo. De novo clustering of long reads by gene from transcriptomics data. *Nucleic Acids Research*, 47(1) :2, 1 2019.

[50] Elaine R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1) :387–402, 9 2008.

[51] Josip Marić, Ivan Sović, Krešimir Križanović, Niranjan Nagarajan, and Mile Šikić. Graphmap2 - splice-aware RNA-seq mapper for long reads. *bioRxiv*, page 720458, 7 2019.

[52] Arfa Mehmood, Asta Laiho, Mikko S Venäläinen, Aidan J McGlinchey, Ning Wang, and Laura L Elo. Systematic evaluation of differential splicing tools for RNA-seq studies. *Briefings in Bioinformatics*, 21(6) :2052–2065, 12 2020.

[53] Christian Mertes, Ines F. Scheller, Vicente A. Yépez, Muhammed H. Çelik, Yingjiqiong Liang, Laura S. Kremer, Mirjana Gusic, Holger Prokisch, and Julien Gagneur. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nature Communications*, 12(1) :1–13, 12 2021.

[54] Michael L. Metzker. Sequencing technologies the next generation, 1 2010.

[55] Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, and Arnaud Lefebvre. CONSENT : Scalable self-correction of long reads with multiple sequence alignment. *bioRxiv*, pages 1–9, 5 2019.

[56] Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Detecting superbubbles in assembly graphs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8126 LNBI, pages 338–348. Springer, Berlin, Heidelberg, 2013.

[57] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12) :1413–1415, 12 2008.

[58] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3) :290–295, 2 2015.

[59] Joseph K. Pickrell, Athma A. Pai, Yoav Gilad, Jonathan K. Pritchard, and J Pfiffner. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genetics*, 6(12) :e1001236, 12 2010.

[60] Fernando Pozo, Laura Martinez-Gomez, Thomas A Walsh, José Manuel Rodriguez, Tomas Di Domenico, Federico Abascal, Jesús Vazquez, and Michael L Tress. Assessing the functional relevance of splice isoforms. *NAR Genomics and Bioinformatics*, 3(2), 4 2021.

[61] Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. High-Throughput Sequencing Technologies, 5 2015.

[62] B L Robberson, G J Cote, and S M Berget. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Molecular and Cellular Biology*, 10(1) :84–94, 1 1990.

[63] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D. Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q. Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S. Butterfield, Richard Newsome, Simon K. Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna Liisa Prabhu, Angela Tam, Yongjun Zhao, Richard A. Moore, Martin Hirst, Marco A. Marra, Steven J.M. Jones, Pamela A. Hoodless, and Inanc Birol. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11) :909–912, 11 2010.

[64] Gustavo A T Sacomoto, Janice Kielbassa, Rayan Chikhi, Raluca Uricaru, Pavlos Antoniou, Marie-France Sagot, Pierre Peterlongo, and Vincent Lacroix. KIS SPLICE : de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, 13(6) :1–12, 2012.

[65] Leena Salmela and Eric Rivals. LoRDEC : Accurate and efficient long read error correction. *Bioinformatics*, 30(24) :3506–3514, 12 2014.

[66] Leena Salmela, Riku Walve, Eric Rivals, and Esko Ukkonen. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, 33(6) :btw321, 6 2016.

[67] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A.

Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage ϕx174 DNA. *Nature*, 265(5596) :687–695, 1977.

[68] Dani Satyawan, Moon Young Kim, and Suk Ha Lee. Stochastic alternative splicing is prevalent in mungbean (Vigna radiata). *Plant Biotechnology Journal*, 15(2) :174–182, 2 2017.

[69] Baptiste Saudemont, Alexandra Popa, Joanna L. Parmley, Vincent Rocher, Corinne Blugeon, Anamaria Necsulea, Eric Meyer, and Laurent Duret. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biology*, 18(1) :208, 12 2017.

[70] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases : Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8) :1086–1092, 2012.

[71] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Josep F Abril, Martin Akerman, Tyler Alioto, Giovanna Ambrosini, Stylianos E Antonarakis, Jonas Behr, Paul Bertone, Regina Bohnert, Philipp Bucher, Nicole Cloonan, Thomas Derrien, Sarah Djebali, Jiang Du, Sandrine Dudoit, Pär G Engström, Mark Gerstein, Thomas R Gingeras, David Gonzalez, Sean M Grimmond, Roderic Guigó, Lukas Habegger, Jennifer Harrow, Tim J Hubbard, Christian Iseli, Géraldine Jean, André Kahles, Felix Kokocinski, Julien Lagarde, Jing Leng, Gregory Lefebvre, Suzanna Lewis, Ali Mortazavi, Peter Niermann, Gunnar Rätsch, Alexandre Reymond, Paolo Ribeca, Hugues Richard, Jacques Rougemont, Joel Rozowsky, Michael Sammeth, Andrea Sboner, Marcel H Schulz, Steven M J Searle, Naryttza Diaz Solorzano, Victor Solovyev, Mario Stanke, Tamara Steijger, Brian J Stevenson, Heinz Stockinger, Armand Valsesia, David Weese, Simon White, Barbara J Wold, Jie Wu, Thomas D Wu, Georg Zeller, Daniel Zerbino, Michael Q Zhang, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12) :1177–1184, 11 2013.

[72] Timothy Sterne-Weiler, Robert J. Weatheritt, Andrew J. Best, Kevin C.H. Ha, and Benjamin J. Blencowe. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell*, 72(1) :187–200, 2018.

[73] Daniel H. Stoloff and Meni Wanunu. Recent trends in nanopores for biotechnology, 8 2013.

[74] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1) :46–53, 1 2013.

[75] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat : discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9) :1105–1111, 5 2009.

[76] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3) :562–78, 2012.

[77] Michael L. Tress, Federico Abascal, and Alfonso Valencia. Alternative Splicing May Not Be the Key to Proteome Complexity, 2 2017.

[78] Robin Lee Troskie, Yohaann Jafrani, Tim R. Mercer, Adam D. Ewing, Geoffrey J. Faulkner, and Seth W. Cheetham. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biology*, 22(1), dec 2021.

[79] Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R Gazzara, Juan Gonzalez-Vallinas, Nicholas F Lahens, John B Hogenesch, Kristen W Lynch, and Yoseph Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5, 2 2016.

[80] John K. Vuong, Chia Ho Lin, Min Zhang, Liang Chen, Douglas L. Black, and Sika Zheng. PTBP1 and PTBP2 Serve Both Specific and Redundant Functions in Neuronal Pre-mRNA Splicing. *Cell Reports*, 17(10) :2766–2775, 12 2016.

[81] Markus C. Wahl, Cindy L. Will, and Reinhard Lührmann. The Spliceosome : Design Principles of a Dynamic RNP Machine, 2 2009.

[82] J. D. Watson and F. H.C. Crick. Molecular structure of nucleic acids : A structure for deoxyribose nucleic acid. *Nature*, 171(4356) :737–738, 1953.

[83] Gene Yeo and Christopher B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In *Journal of Computational Biology*, volume 11, pages 377–394. Mary Ann Liebert, Inc., 7 2004.