Développement d'outils biostatisques et bioinformatiques de prédiction et d'analyse des défauts de l'épissage : application aux gènes de prédisposition aux cancers du sein et de l'ovaire

Raphaël Leman

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité RECHERCHE CLINIQUE, INNOVATION TECHNOLOGIQUE, SANTE PUBLIQUE

Préparée au sein de l'Université de Caen Normandie

Développement d'outils biostatisques et bioinformatiques de prédiction et d'analyse des défauts de l'épissage : application aux gènes de prédisposition aux cancers du sein et de l'ovaire

Présentée et soutenue par
Raphael LEMAN

Thèse soutenue publiquement le 13/12/2019 devant le jury composé de		
Mme MARIE PIERRE BUISINE	Professeur des universités, 59 CHRU de LILLE	Rapporteur du jury
Mme FABIENNE LESUEUR	Directeur de recherche, Institut Curie Paris	Rapporteur du jury
M. CLAUDE HOUDAYER	Professeur des universités, 76 CHU de ROUEN	Membre du jury
M. JEAN MULLER	Maître de conférences, Université de Strasbourg	Membre du jury
M. NICOLAS SEVENET	Professeur des universités, Université de Bordeaux	Président du jury
Mme ALEXANDRA MARTINS	Directeur de recherche, Université Rouen Normandie	Directeur de thèse
Mme SOPHIE KRIEGER	Maître de conférences HDR, Université Caen Normandie	Co-directeur de thèse

Thèse dirigée par ALEXANDRA MARTINS et SOPHIE KRIEGER, Génomique et médecine personnalisée du cancer et troubles neurologiques

UNIVERSITÉ CAEN NORMANDIE

ED 497 NBISE
Normande de Biologie Intégrative, Santé, Environnement

Centre de Lutte contre le Cancer François Baclesse
unicancer NORMANDIE - CAEN

13/12/2019

Développement d'outils biostatistiques et bioinformatiques de prédiction et d'analyse des défauts de l'épissage : application aux gènes de prédisposition aux cancers du sein et de l'ovaire.

Travaux encadrés par le Dr Sophie KRIEGER (MCU-PH) et le Dr Alexandra MARTINS (DR)

Dr Raphaël LEMAN
LABORATOIRE DE BIOLOGIE ET DE GENETIQUE DU CANCER, INSERM U1245

Résumé

L'analyse des défauts d'épissage est particulièrement complexe. Outre la diversité des transcrits présents à l'état physiologique, les variations nucléotidiques peuvent induire des modifications hétéroclites de l'épissage. Ces variations, appelées variants splicéogéniques, et leur impact au niveau de l'épissage, sont à même de modifier plus ou moins sévèrement le phénotype de l'individu.

Au cours de ce travail de thèse, nous nous sommes intéressés à trois grands aspects de l'étude des défauts de l'épissage : (i) la prédiction de ces défauts d'épissage, (ii) l'analyse des données de RNA-seq et (iii) le rôle de l'épissage dans l'interprétation de la pathogénicité d'un variant pour la prédisposition aux cancers du sein et de l'ovaire (syndrome HBOC).

Nous avons optimisé les recommandations en vigueur pour identifier les variants splicéogéniques au sein des séquences consensus des sites d'épissage. Ce travail a conduit à la publication d'un nouvel outil SPiCE (*Splicing Prediction in Consensus Elements*), développé sur 395 variants. SPiCE a le potentiel d'être une aide à la décision pour guider les généticiens vers ces variants splicéogéniques, grâce à une exactitude de 94.4 %. Puis, nous avons comparé les outils de prédiction des points de branchement. Pour cela, une collection sans précédente de 120 variants avec leurs études ARN a été établi dans la région des points de branchements. Nous avons ainsi révélé que ces outils de prédictions sont aptes à prioriser les variants pour des études ARN dans ces régions jusque-là peu étudiées. Pour étendre les prédictions des variants splicéogéniques au-delà d'un motif spécifique, nous avons construit l'outil SPiP (*Splicing Prediction Pipeline*). SPiP utilise un ensemble d'outils pour prédire un défaut d'épissage quel que soit la position du variant. Ainsi, SPiP peut ainsi s'adresser à la diversité des défauts d'épissage avec une exactitude de 80.21 %, sur une collection de 2 784 variants.

Les données issues du RNA-seq sont complexes à analyser, car il existe peu d'outils pour annoter finement les épissages alternatifs. Aussi nous avons publié l'outil SpliceLauncher. Cet outil permet de déterminer une grande diversité de jonctions d'épissage, indépendamment des systèmes RNA-seq utilisés. Cet outil renvoie aussi les résultats sous formes graphiques pour faciliter leur interprétation.

Puis nous avons évalué le rôle de l'épissage alternative dans l'interprétation à usage clinique d'un variant. Le gène *PALB2*, impliqué dans le syndrome HBOC, a été utilisé comme modèle d'étude. Nous avons ainsi démontré que l'épissage alternatif de *PALB2* est apte à remettre en cause la pathogénicité de certains variants. La collecte de données fonctionnelles et cliniques sont donc nécessaires pour conclure sur leur pathogénicité.

Nos travaux illustrent ainsi l'importance de la caractérisation et de l'interprétation des modifications de l'épissage pour répondre aux défis présents et futurs du diagnostic moléculaire en génétique.

Mots-clés : épissage, variants, syndrome HBOC, prédiction, RNA-seq, SPiP, SPiCE, SpliceLauncher

Abstract

Analysis of splicing defects is particularly complex. In addition to the diversity of physiological transcripts, nucleotidic variations can induce heterogeneous alteration of splicing. These variations, called spliceogenic variants, and their impact on splicing, can involve severe consequences on the individual phenotype.

In this thesis work, we focused on three main aspects of the study of splicing defects: (i) the prediction of these splicing defects, (ii) the analysis of RNA-seq data and (iii) the role of splicing in interpreting the pathogenicity of a variant for the hereditary breast and ovarian cancers (HBOC syndrome).

We optimized the current recommendations to identify spliceogenic variants within the consensus sequences of splicing sites. This work led to the publication of a new tool, SPiCE (Splicing Prediction in Consensus Elements), developed on 395 variants. SPiCE has the potential to be a decision support tool to guide geneticists towards these spliceogenic variants, with an accuracy of 94.4%. Then, we compared the tools dedicated to branch points prediction. For this purpose, an unprecedented collection of 120 variants with their RNA studies has been established in the branch point region. Thus, we revealed these prediction tools are able to prioritize variants for RNA studies in these hitherto poorly studied regions. To extend the predictions of spliceogenic variants beyond a specific motif, we built SPiP (Splicing Prediction Pipeline) tool. SPiP uses a set of tools to predict a splicing defect regardless of the variant position. Thus, SPiP can address the diversity of splicing defects with an accuracy of 80.21%, on a collection of 2,784 variants.

The data from the RNA-seq are complex to analyze, as there are few tools to finely annotate alternative splices. Also we published SpliceLauncher tool. This tool allows to determine a wide variety of splicing junctions, independently of RNA-seq systems used. This tool also returns the results in graphical form to make interpretation user-friendly.

Then we evaluated the role of alternative splicing in the clinical interpretation of a variant. The *PALB2* gene, involved in HBOC syndrome, was used as a study model. Thus, we demonstrated that the alternative splicing of *PALB2* is able of challenging the pathogenicity of certain variants. Collection of functional and clinical data is therefore necessary to conclude on their pathogenicity.

Our work thus illustrates the importance of characterizing and interpreting splicing modifications to meet the current and future challenges of molecular diagnosis in human genetics.

Keywords: splicing, variants, HBOC syndrome, prediction, RNA-seq, SPiP, SPiCE, SpliceLauncher

Remerciements

Nous remercions les membres du jury, tout particulièrement les rapporteurs, le Professeur Marie-Pierre Buisine et le Docteur Fabienne Lesueur, pour avoir acceptés de s'immerger dans ce sujet complexe mais au combien passionnant des défauts d'épissage. Nous remercions également le Professeur Nicolas Sevenet et le Docteur Jean Muller d'examiner ce travail au regard de leurs expertises.

Nous remercions également le Professeur Claude Houdayer dont son implication est impossible à estimer tant sa participation à ce travail a été enthousiaste et cruciale.

Nous sommes reconnaissant envers le Professeur Thierry Frébourg, directeur de l'unité Inserm U1245, et le Docteur Dominique Vaur, Directeur du laboratoire de biologie et de génétique du cancer du Centre François Baclesse, pour nous avoir offert les infrastructures nécessaires au déroulement de cette thèse.

Nous remercions également nos collègues biologistes le Docteur Laurent Castera, le Docteur Etienne Mueller et le Docteur Agathe Ricou pour leurs conseils avisés et pour nous avoir partagé leur expérience dans l'interprétation des variants.

Nous congratulons également le Docteur Nicolas Goardon pour sa veille bibliographique et Angelina Legros pour son assistance technique.

Nous sommes également reconnaissants envers nos collègues bioinformaticiens (Docteur Alexandre Atkinson, Baptiste Brault, Thibaut Lavole, Germain Paimparay et Antoine Rousselain) pour leur avis d'expert dans la construction de nos outils bioinformatiques.

Nous remercions le Docteur Pascaline Gaildrat et le Docteur Alexandra Martins pour leur participation à ce travail de thèse.

Nous félicitons Laetitia Meulemans, le Docteur Omar Soukarieh, et le Docteur Hélène Tubeuf pour leurs travaux innovant dans l'étude et l'interprétation des défauts d'épissage.

Nous sommes débiteurs pour le Docteur Sabine Raad et le Docteur Isabelle Tournier pour nous avoir partagé leurs données RNA-seq.

Nous remercions également Valentin Harter et le Professeur Jean-Philippe Vert, nos oasis statistiques dans un monde de biologistes.

Nous sommes également reconnaissants envers le Docteur Laurent Poulain et les membres de son équipe BioTICLA de l'unité Inserm U1199 ANTICIPE, pour nous avoir offert leurs assistances techniques.

Nous complimentons aussi les membres du réseau épissage de GGC ainsi que les membres de l'unité Inserm UMR1078, ceux du service de Génétique et Biologie Moléculaires de l'HUPC Hôpital Cochin

et les membres du laboratoire de Génétique du GH Saint-Louis-Lariboisière-Fernand Widal, pour leur participation. A l'instar de rivières devenant des fleuves puis des océans, leurs efforts acharnés et continus, pour caractériser les défauts d'épissage variant par variant, ont magnifié ce travail de thèse.

Nous sommes également redevables envers le Docteur Amanda Spurdle et les membres du consortium ENIGMA et tout particulièrement le Docteur Miguel de la Hoya et le Docteur Logan Walker.

Agradecemos a Miguel de la Hoya por asociarnos al estudio del gen PALB2.

We would like to thank Logan Walker to associate us at the QC RNA-seq Project.

Wij feliciteren Rien Blok met zijn studie over alternatieve verbindingen van RAD51C/D genen.

Také děkujeme Petra Kleiblova za pomoc v projektu QC RNA-seq.

Vi gratulerar också Anders Kvist med hans innovativa tillvägagångssätt.

Vi takker også Thomas van Overeem Hansen for at dele hans RNA-data.

Je suis également reconnaissant envers le Docteur Sophie Krieger et le Docteur Alexandra Martins pour avoir encadré ce travail de thèse.

Je tiens aussi personnellement à remercier Sophie Krieger pour avoir pris le risque d'engager un « étranger » pour porter son projet de recherche. Et je m'excuse pour toutes les séances de supplices statisticiennes endurées. Mais grâce à ce projet, j'ai découvert tout un univers que je ne connaissais guère.

Je remercie aussi mes parents, mes supporteurs de l'ombre qui sont toujours là pour partager les joies et les peines. Et aussi n'oublions pas les toutous, Lana, Lilou et Nasca. D'ailleurs, Nasca adore courir sur les plages après les mouettes. Donc merci pour lui, Sophie de m'avoir fait venir en Normandie.

Je remercie aussi Manu du Centre Régional de Tir de Bretteville sur Odon, pour m'avoir fait découvrir un sport bien souvent méconnu.

Table des matières

LISTE DES FIGURES

LISTE DES TABLEAUX

INDEX DES ABBREVIATIONS

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

INTRODUCTION

L'analyse des défauts d'épissage peut se révéler particulièrement complexe. Outre la diversité des transcrits présents à l'état physiologique, les variations nucléotidiques peuvent induire des modifications complexes de l'épissage. Aussi dans cette partie introductive nous faisons l'état des connaissances des mécanismes d'épissage, ainsi que des moyens mis en œuvres pour l'étudier, tant d'un point de vue fonctionnel que prédictif. De plus nous avons pris comme modèle d'étude la prédisposition aux cancers du sein et de l'ovaire pour approfondir la relation entre un défaut d'épissage et un phénotype clinique.

I. L'épissage

1. L'épissage : étape clé dans la maturation des ARN pré-messagers

Les acides ribonucléiques (ARN), aussi appelés *ribonucleic acids* (RNA), jouent de nombreux rôles cruciaux chez les eucaryotes. Les molécules d'ARN peuvent être regroupées en deux grands types : les ARN codants ou ARN messagers (ARNm), *messenger RNA* (mRNA), et les ARN non codants (ARNnc). Bien que les ARNm ne représentent en moyenne que 1 % des molécules d'ARN, les ARNm constituent le support de l'information entre la séquence génique et protéique lors de l'étape de traduction (Figure 1). Parmi les ARNnc, les ARN ribosomaux et de transferts jouent principalement un rôle de partenaire dans la traduction protéique et sont synthétisés par les ARN polymérases I. Les petits ARN nucléaires, ou snRNA, sont impliqués dans la régulation des modifications post-transcriptionnelles de l'ARNm dont l'épissage [1]. D'autres ARNnc participent à la régulation de l'expression génique. Parmi eux, les micro ARN ciblent la dégradation des ARNm en se liant spécifiquement à la partie 3' non traduite des ARNm [2]. Les longs ARN non codants régulent eux de manière plus complexe cette expression [3].

Depuis leur synthèse par les ARN polymérases II, les ARNm subissent plusieurs étapes de maturation pour pouvoir être éligibles à la traduction en protéine : l'ajout d'une coiffe 7-methylguanosine en 5', la polyadénylation en 3' et enfin l'épissage. Si les deux premières étapes sont indépendantes de la séquence de l'ARN pré-messager (pré-ARNm), dans leur processus, l'épissage est étroitement lié à la séquence de la molécule. L'épissage a été décrit pour la première fois en 1977 par l'équipe de Richard J. Roberts en utilisant l'adénovirus comme modèle [4]. Il consiste en l'assemblage des séquences codantes du pré-ARNm, les séquences non codantes, entre celles codantes, étant excisées de la molécule. Les séquences codantes et non codantes sont nommées respectivement exons et introns [5]. Les exons ont une longueur médiane de 133 nucléotides (nt) tandis que les introns ont une taille médiane de 1 851 nt, calculée à partir de la base de données RefSeq [6]. Ceci illustre que moins de 10 % de la molécule pré-ARNm est codante. En conséquence l'épissage est une étape majeure de la maturation des ARNm.

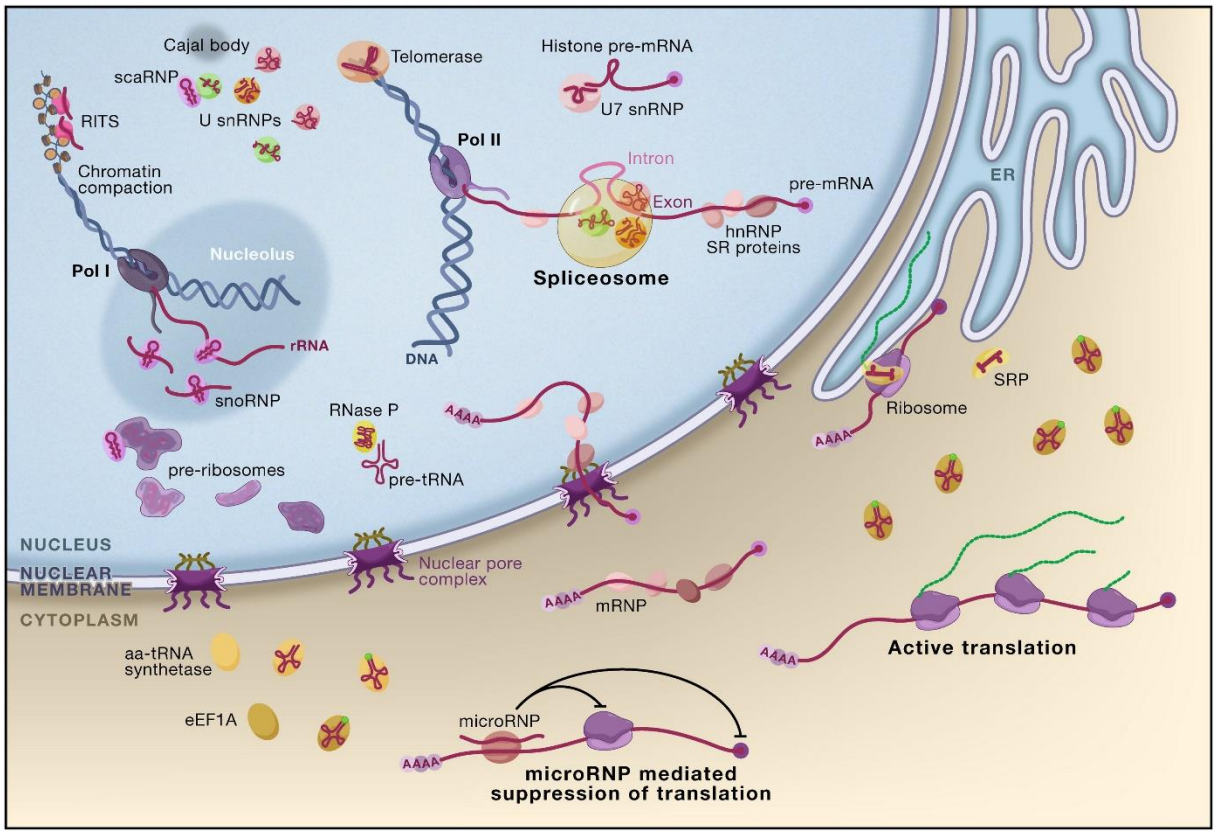


Figure 1 : Rôles des ARN dans l'expression des gènes (d'après [7]). Ici la majorité des ARNs et des complexes ribo-nucléoprotéiniques (RNPs) sont dépeintes. Suite à la transcription par l'ARN polymérase II (RNA Pol II), les ARN pré-messagers (pre-ARNms) sont liés par diverse protéines dont les protéines *serine-arginine-rich* (hnRNP et SR). Les pre-ARNms, avec les exons (en rouge) et les introns (en rose), sont traités par le splicéosome. Certains ARNs tels que les ARN de pré-transferts et les ARNm codant pour les histones, sont aussi traités par des RNPs spécifique (RNase P et U7 snRNP). Les small nucleolar RNPs (snoRNPs) et small Cajal body RNPs (scaRNPs) médient la maturation des ARN utilisés par les RNPs tels que les ARN ribosomaux (produits par l'ARN polymérase I ou RNA Pol I) et les petits ARN nucléaires (snRNAs). Les petits ARN peuvent aussi former des microARN (microRNPs) qui régulent la traduction des transcrits. Dans le cytoplasme, le ribosome représente la RNP clé qui dirige la traduction des ARNm en protéine. Il s'associe aussi avec la protéine *signal recognition particle* (SRP) pour permettre la translocation de la protéine dans le réticulum endoplasmique (ER).

a. La machinerie d'épissage : le splicéosome

L'épissage des ARNm est assuré par un complexe protéique nommé splicéosome. Le splicéosome est composé de complexes ribo-nucléoprotéiniques (RNPs). Les principales RNPs sont les U1, U2, U4, U5 et U6 [8]. En complément, le splicéosome est assisté par un ensemble d'autres RNPs dont SAP155, $U2AF^{65}$, $U2AF^{35}$, et des protéines riches en sérine et arginine *Serin-Arginin rich protein* (protéines SR). La principale fonction de ces dernières est la reconnaissance des motifs d'épissage présents dans le pré-ARNm. En effet le splicéosome utilise des motifs hautement conservés, dits canoniques, pour définir les jonctions exons/introns. Ces motifs sont situés dans les introns en partie 5' et 3' et sont au nombre de trois. En partie 5' de l'intron est observé majoritairement le site donneur caractérisé majoritairement par un motif canonique GT. Dans 0,82 % des sites donneurs humains le motif est GC [9]. En partie 3'

deux motifs participent à l'épissage, le site accepteur pourvu d'un motif canonique AG et en amont de ce site le point de branchement identifié par une adénosine.

Pour procéder à l'épissage de l'intron, le splicéosome découpe le pré-ARNm au niveau du site donneur. La partie libre 5' de l'intron subit une trans-estherification avec le point de branchement. Cette étape conduit à la formation d'un ARN lasso. Le splicéosome tronque le pré-ARNm à hauteur du site accepteur et associe les deux séquences exoniques qui bordaient l'intron. Le reliquat d'intron, devenu un ARN lasso, est libéré par le splicéosome pour être ensuite dégradé (Figure 2).

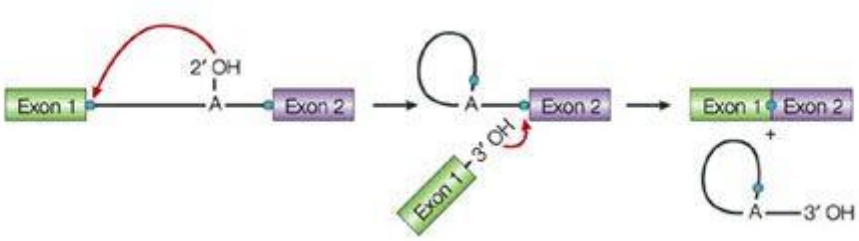


Figure 2 : Représentation schématique de l'épissage d'un intron (d'après [10]). La première et la deuxième étape d'épissage impliquent une attaque nucléophilique (flèche rouge) sur la liaison phosphodiester terminale (points bleus) par le 2'hydroxyle du point de branchement (A) et par le 3' hydroxyle de l'exon en amont. Les exons ligués et l'intron lasso sont représentés à droite.

Un second splicéosome composé des protéines U11, U12, U4, U5 et U6 participe également à l'épissage. Il est nommé le splicéosome U12 mineur par opposition au précèdent splicéosome appelé le splicéosome U2 majeur. En effet, seulement 0.1 % des introns humains sont épissés par ce splicéosome U12 mineur [11]. Si le processus d'épissage est similaire entre les deux splicéosomes, ils différent par les motifs canoniques reconnus du pré-ARNm. Le site donneur est défini par le motif canonique AT et le site accepteur par le motif canonique AC.

Les motifs canoniques présents sur la séquence du pré-ARNm guident le splicéosome pour épisser les introns. Or la séquence de ces motifs étant courte, un problème mathématique se pose rapidement. En effet, à titre d'exemple, le gène *RAD51B* contient une séquence d'environ 776 000 nt. En supposant que la probabilité de trouver un motif canonique donneur ou accepteur soit de $^{1}/_{16}$ (6,25 %), l'espérance mathématique du nombre de sites d'épissage est de 48 515 sites. Mais *RAD51B* ne comprend que 11 exons (NM_133509), soit 20 sites d'épissage. Dès lors il apparait que le splicéosome et les motifs canoniques ne sont pas les seuls acteurs de l'épissage.

b. Les motifs d'épissage

Très tôt après la découverte de l'épissage en 1977, l'équipe de Chambon a montré que la séquence du pré-ARNm autour des sites canoniques joue un rôle majeur dans la reconnaissance des sites d'épissage [12]. Dès lors plusieurs études ont commencé à aligner les séquences des sites canoniques pour définir une séquence consensuelle observée autour de ces sites. L'ensemble de ces résultats a été résumé par l'équipe de Phillip A. Sharp en 1999, aboutissant à le définition des trois séquences consensuelles retrouvées autour des sites canoniques (Figure 3). La séquence des sites donneurs d'épissage implique les six premières bases de l'intron et les trois dernières bases de l'exon. Le motif consensus est VAG|**GT**RAGT, où | est la jonction exon/intron et **GT** le motif canonique. La séquence du site accepteur inclus les deux premières bases de l'exon et les 12 dernières bases de l'intron. La séquence consensus du site accepteur se caractérise aussi par un enrichissement en pyrimidine à partir du 5ème nucléotide dans l'intron, appelé tract polypyrimidique. Tandis que le 4ème nucléotide dans l'intron n'est pas conservé. La séquence retrouvée est YYYYYYYYNC**AG**|GD, où | est la jonction intron/exon et **AG** le site canonique.

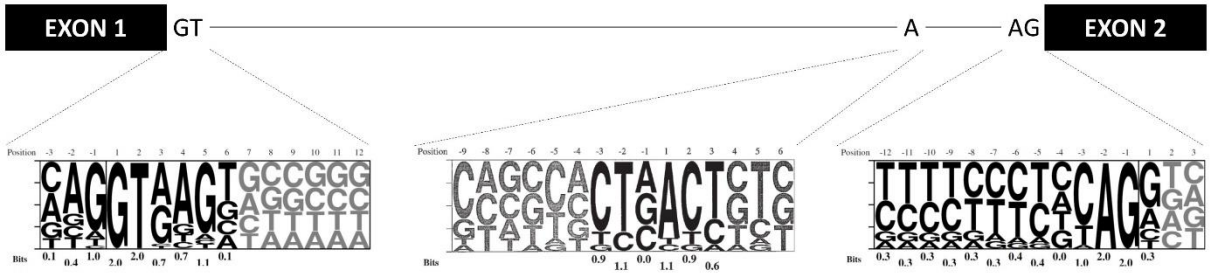


Figure 3 : Séquences consensuelles autour des motifs canoniques d'épissage des introns humains U2 (adaptée de [13]). A chaque position de la séquence, les fréquences f1, f2, f3, et f4 des quatre nucléotides A, C, G, et T de l'ADN sont représentées par la hauteur des lettres correspondantes.

Le point de branchement est représenté par la plus courte séquence consensus (6 nt) avec seulement l'adénosine du point de branchement et une thymidine, deux bases en amont, hautement conservées (C**T**R**A**YY). La difficulté majeure concernant les points de branchement est de connaître leur position exacte en amont du site accepteur. La courte taille des motifs et le fait que l'ARN lasso soit rapidement dégradé a limité l'étude *in vitro* des points de branchement à de la mutagénèse dirigée en amont des sites accepteurs, intron par intron (ex : [14]). Il faudra attendre près de 40 ans depuis la découverte du mécanisme d'épissage en 1977 pour obtenir la première étude à large échelle décrivant expérimentalement les points de branchement [15]. Les auteurs de cette étude ont proposé une approche originale consistant à séquencer à haut débit l'ARN lasso par une reverse-transcription spécifique de la jonction site donneur/point de branchement combinée avec une dégradation de l'ARN linéaire. Cette cartographie a notamment permis de montrer que plus de 95 % des points de branchement humain sont situés entre 44 et 18 nt en amont des sites accepteurs.

Si la séquence consensus des sites d'épissage a été largement étudiée et décrite, il faudra attendre le début des années 2000 pour prouver qu'elles ne sont pas suffisantes pour définir les jonctions exon/intron [16]. En effet, il a été identifié par la suite des motifs autour des sites d'épissage capables de favoriser ou d'inhiber l'utilisation de ces sites. Pour les motifs introniques, ils sont habituellement nommés *Intronic Splicing Enhancers* (ISEs) ou *Silencers* (ISSs). Les motifs exoniques sont eux nommés *Exonic Splicing Enhancers* (ESEs) ou *Silencers* (ESSs). L'ensemble de ces motifs sont regroupés sous l'appellation *Splicing Regulatory Elements* (SREs), et *Exonic Splicing Regulators* (ESRs) pour les motifs exoniques. Ces motifs de taille moyenne comprise entre 4 et 8 nt sont des sites de fixation de complexes RNPs associées au splicéosome.

Les motifs *enhancers* sont associés aux protéines SR, par exemple SC35 et SF2/ASF [17]. Les motifs *silencers* sont identifiés par la classe des protéines hnRNP, incluant entre autre hnRNP I et hnRNP A1 [18]. Si la séquence des sites d'épissage peut-être facilement étudiée par alignement des jonctions intron/exons, l'étude des SREs s'avèrent plus complexe. De nombreuses études par différentes approches *in silico, in vivo* ou combinant des analyses à haut débit *in vitro* ont permis d'étudier la séquence de ces SREs et leur impact sur l'épissage [19]–[25]. Par une approche à haut débit, l'équipe de Ronsenberg a étudié l'influence des 4096 motifs hexamèriques possibles sur la reconnaissance des sites d'épissage [25]. Il en a résulté que 82.9 % des (3 396/4 096) motifs possibles étaient significativement associés à l'utilisation des sites d'épissage, leur association étant plus importante au niveau exonique. Il en résulte que c'est autant un environnement de SREs plutôt qu'un seul motif qui permet la reconnaissance des sites d'épissage [26].

2. L'épissage alternatif

La multiplicité des motifs d'épissage tout au long de la séquence pré-ARNm indique que l'épissage ne conduit pas nécessairement à une même molécule ARNm. En effet l'étude de l'expression des gènes a révélé qu'un même pré-ARNm pouvait donner un grand nombre d'ARNm différents. Un exemple particulièrement impressionnant est le gène *Dscam* de *Drosophila melanogastar* qui produit 38 016 ARNm différents [27] à partir de la même molécule pré-ARNm. Ce phénomène appelé épissage alternatif est également retrouvé chez l'homme. A titre d'exemple, le gène *KCNMA1* génère plus de 500 isoformes d'ARNm [28], [29]. Outre le grand nombre d'isoformes ARNm générées par ce procédé, l'épissage alternatif est un processus particulièrement commun chez l'homme. Il concerne plus de 90 % des gènes multi-exons [30]. Cette constante diversité transcriptionnelle a joué un rôle crucial dans le processus d'évolution des vertébrés [31]. De plus chaque molécule ARNm produite conduit à la genèse de protéines dont les rôles fonctionnels peuvent être différents voire opposés. Le gène *Bcl-x* illustre cette dichotomie puisque le transcrit pleine longueur génère une protéine impliquée dans l'activation de l'apoptose, alors qu'il existe un transcrit alternatif, plus court, produisant une protéine inhibant

l'apoptose [32]. L'épissage alternatif est également soumis à une régulation très fine. Par exemple le même gène peut produire différents transcrits au cours du développement d'un individu [33]. L'épissage alternatif est également tissu-spécifique. Le tissu nerveux ayant notamment un profil d'expression des gènes divergeant par rapport au reste des tissus [34]. Il a également été montré que l'épissage alternatif peut être sollicité en réponse à un stimulus extérieur [35].

Pour obtenir cette multiplicité en ARNm à partir d'une seule séquence de pré-ARNm, l'épissage alternatif implique 3 grands mécanismes d'épissage (Figure 4):

- Le saut d'exon.
- L'utilisation de nouveaux sites d'épissage.
- Le non-épissage de l'intron.

Le saut d'exon affecte les exons cassettes, c'est-à-dire les exons bordés par des régions introniques. La perte d'exon peut soit affecter un seul exon ou soit plusieurs exons consécutifs.

L'utilisation d'un nouveau site d'épissage peut concerner soit un site donneur ou accepteur. Ce site d'épissage peut être intronique ou exonique, par rapport au transcrit de référence, entraînant soit l'inclusion d'une partie d'intron ou soit la délétion d'une partie d'exon. Deux nouveaux sites peuvent se combiner dans l'intron pour créer un nouvel exon, appelé pseudo-exon. Cette combinaison peut aussi apparaître dans un exon. Ceci a pour conséquence de mimer un intron au sein de l'exon et donc d'entraîner une perte partielle de cet exon.

Le non épissage de l'intron correspond à la rétention complète de l'intron dans l'ARNm. Cet événement souvent considéré comme minime, s'avère être prépondérant dans certains processus physiopathologiques comme la tumorigénèse [36]. Ces mécanismes d'épissage alternatif peuvent également apparaître de façon conjointe.

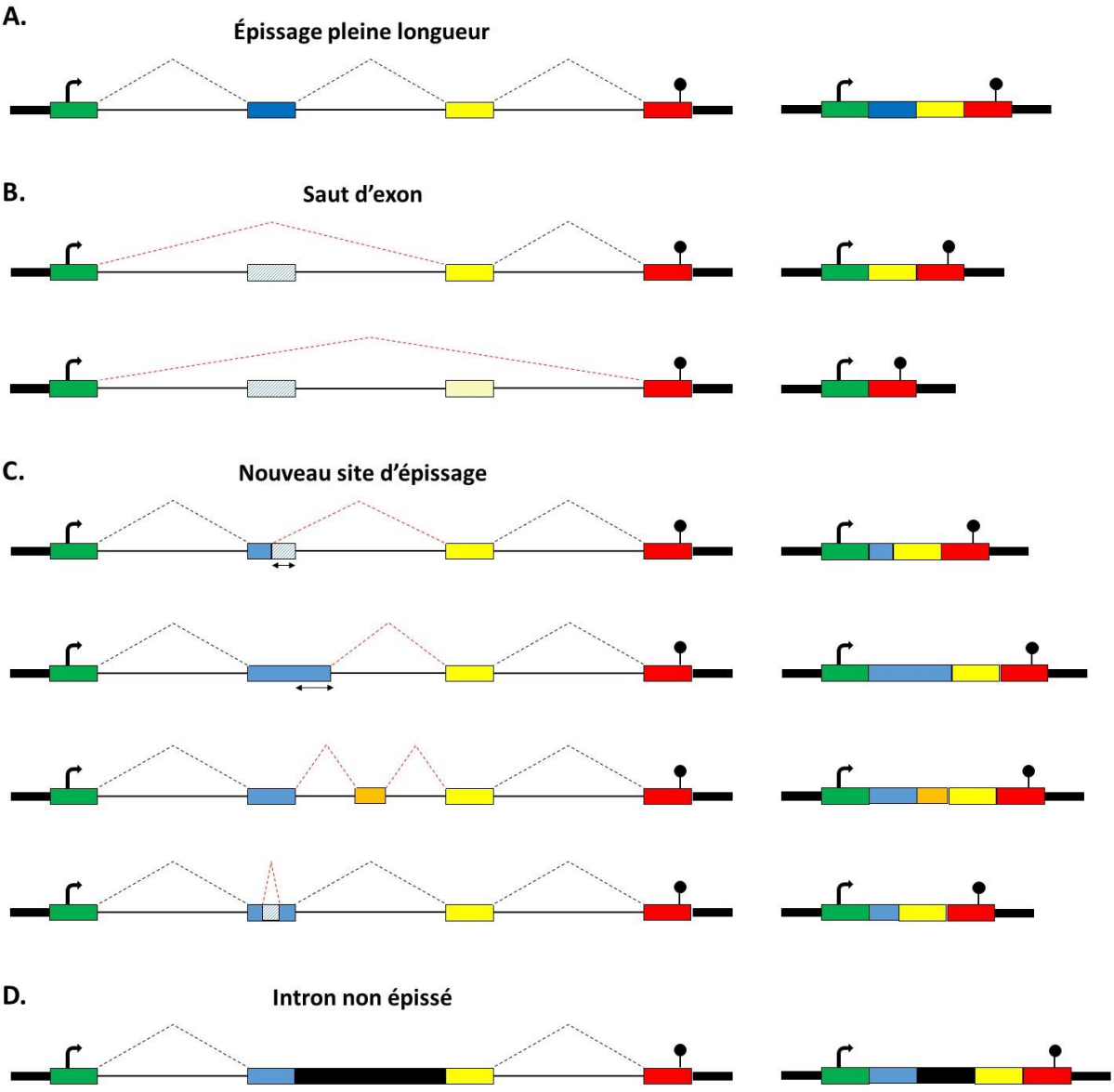


Figure 4 : Différents mécanismes impliqués dans l'épissage alternatif. **A.** Une représentation schématique d'un pré-ARNm avec 4 exons (rectangles) et 3 introns (lignes noires), le codon d'initiation est représenté par une flèche incurvé et le codon stop par un point. Les jonctions d'épissage sont affichées en pointillés et l'ARNm après épissage est figuré à droite. **B.** Saut d'exon soit simple ou soit multiple. **C.** Utilisation d'un ou plusieurs nouveaux sites d'épissage. Si le nouveau site est exonique alors l'ARNm sera délesté d'une partie de l'exon, ou si le nouveau site est intronique alors l'ARNm comprendra une partie de l'intron. Les nouveaux sites d'épissage peuvent être utilisés par paires avec soit la création d'un exon s'ils sont introniques ou soit la création d'un intron s'ils sont exoniques. **D.** Ici l'intron n'est pas reconnu comme tel par le splicéosome et donc la séquence complète de l'intron est retrouvée dans l'ARNm.

Les modifications apportées aux ARNm peuvent impacter le cadre de lecture avec la création de codons stops prématurés ou PTCs, *Premature Termination Codons*. Les PTCs peuvent conduire à la traduction de protéines tronquées avec un potentiel impact délétère par perte de fonction ou bien par effet dominant négatif.

La cellule dispose d'un système nommé NMD pour *Nonsense-Mediated Decay*, permettant notamment de contrôler l'apparition d'ARNm doté de PTCs. Le NMD intervient lors du premier cycle de traduction de l'ARNm par le ribosome. Trois facteurs protéiques, nommés *up-frameshift protein* (UPF) jouent un rôle clé dans l'activation du NMD. Notamment UPF1, une fois activé, conduit à la dégradation de l'ARNm. Les protéines UPF2 et UPF3 sont associées à un plus grand complexe protéique ciblant les jonctions exons/exons, appelé *exon-junction complex* (EJC). Ce complexe EJC est recruté par le splicéosome pour être fixé entre 20-24 nt en amont des jonctions exon/exon. Il correspond ainsi à une « empreinte » laissée par l'épissage pour marquer les points de cassure du pré-ARNm [37]. En effet, la plupart des codons stop naturels sont présents sur le dernier exon. La plupart des codons stops identifiés en amont de jonction exon/exon peuvent être considérés comme des PTC [38]. Une fois ce marquage EJC effectué lors du premier cycle de traduction dans le cytosol, si le ribosome ne détecte aucun codon stop avant de rencontrer le complexe EJC, alors le ribosome dissocie ce complexe et poursuit la traduction. Si le ribosome arrive à hauteur d'un codon stop en l'absence de complexe EJC alors la traduction s'arrête en libérant le polypeptide et le cycle de traduction recommence. Si le ribosome rencontre un PTC en amont d'un complexe EJC, alors le ribosome active les facteurs eRFs (*eukaryotic Release Factors*). Ces derniers permettent l'association de UPF1 avec les deux protéines UPF2 et UPF3, déjà présentes sur l'EJC. L'association des 3 protéines UPFs entrainent la dégradation de l'ARNm (Figure 5). Etant donné l'encombrement stérique du complexe EJC, les PTCs situés à moins de 50 nt d'un site donneur ne sont pas détectés par le NMD [39], [40]. Il a aussi été récemment montré que l'activation de UPF1, et donc du NMD, peut dépendre de la distance entre un codon stop et la partie 3' polyadénylée [41].

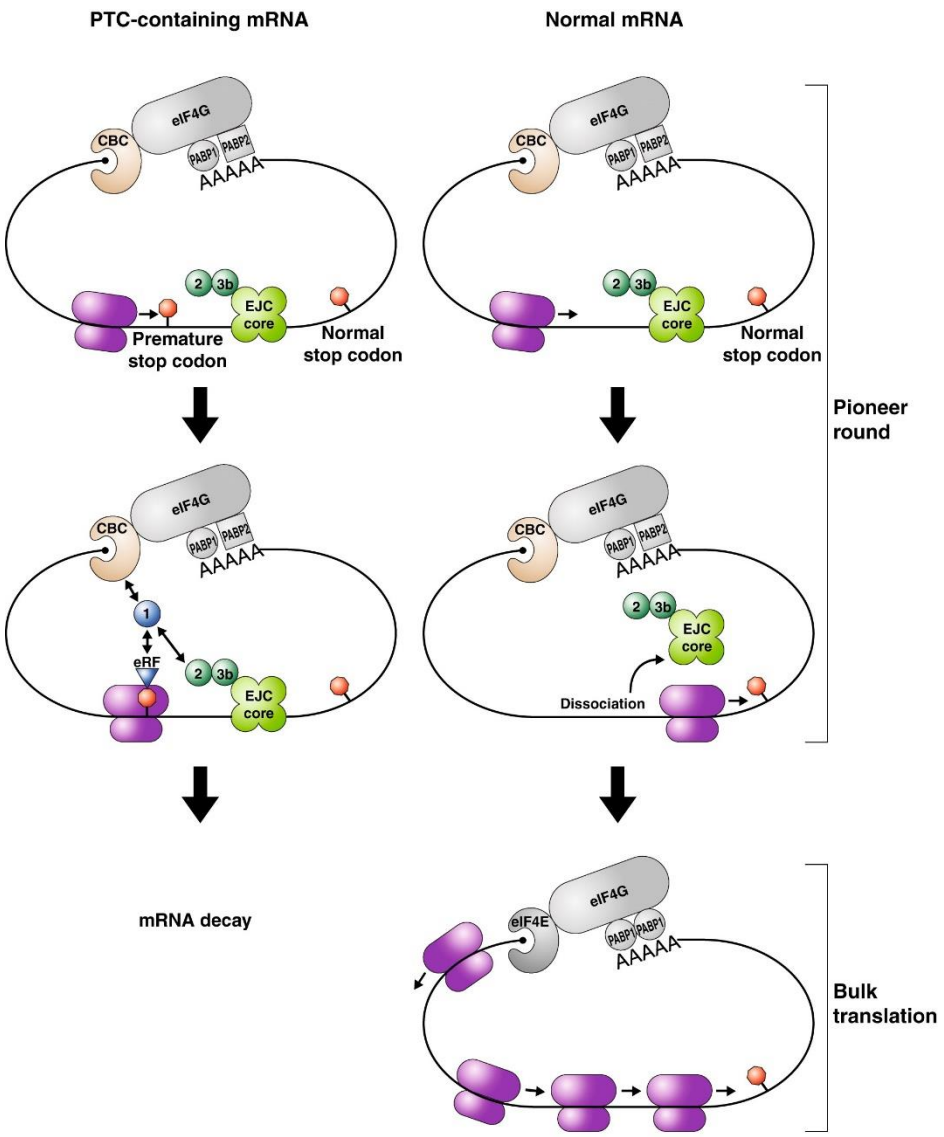


Figure 5 : Voie de signalisation du Nonsense-Mediated Decay (NMD) qui survient lors de l'identification d'un PTC durant un unique tour de traduction « *pioneer* » (d'après [42]). Le NMD est spécifiquement actif quand UPF1(up-frameshift protein 1) interagit avec UPF2, qui se lie avec UPF3b. UPF2 et UPF3b font parties de l'EJC, qui est recruté au niveau des jonctions exon-exon durant l'épissage des ARNm. Dans le cas d'un transcrit aberrent (panel de gauche), il y a typiquement au moins un EJC déposé en aval d'un codon stop prématuré, lui permettant d'interagir avec UPF1 recruté par eRF1 et eRF3. L'interaction entre UPF1 et UPF2 est aussi favorisée par le CBC, qui est lié à la partie 5' de l'ARNm durant le tour *pioneer* de traduction. A l'inverse, un transcrit normal (panel de droite) échappe au NMD car tous les EJC (seulement un est figuré sur ce schéma) sont en amont du codon stop et sont ensuite éliminés par le ribosome avant que UPF1 soit recruté. Après ce premier tour de traduction, les transcrits normaux se voient dotés de nouvelles protéines en 5' et procèdent à la traduction en masse de la protéine.

3. Des variants génétiques aux défauts d'épissage

Seulement quelques années après la découverte de l'épissage, il a été montré que des altérations génétiques, appelées variants, pouvaient impacter ce dernier. Ce phénomène a été pour la première fois décrit dans le cadre des β-thalassémies [43]. En effet ces variants affectant l'épissage, ou variants splicéogéniques, ont été montrés comme étant une cause de cette maladie génétique. Dès lors, de nombreuses études ont investigué les variants capables de modifier l'épissage et par quels mécanismes.

Les variants splicéogéniques peuvent affecter l'épissage soit en agissant en *cis* ou soit en *trans*. Les variants splicéogéniques *trans* sont ceux modifiant l'épissage en affectant les facteurs protéiques impliquées dans l'épissage des ARNm [44]. Les variants *cis* regroupent ceux altérant les motifs d'épissage présents sur la molécule pré-ARNm du gène concerné. Les variants splicéogéniques *trans* sont relativement peu représentés par rapport aux variants *cis*. Une des principales raisons est que les variants *trans* sont généralement létaux durant le développement embryonnaire [45].

A l'inverse, les variants *cis* semblent être largement représentés. En effet, ces derniers peuvent être observés pour n'importe quel gène codant. Aussi dès le début des années 2000, il a été émis l'hypothèse que les variants splicéogéniques, loin d'être un phénomène isolé, seraient la cause la plus fréquente de maladie héréditaire [46]. Cette hypothèse semble se confirmer au regard des bases de données de variants. A titre d'exemple, la *Human Genome Mutation Database* (HGMD) (http://www.hgmd.cf.ac.uk/ac/index.php), ayant pour but de collecter des variants associés à des pathologies héréditaires humaines, rapporte 15 % de variants situés dans les sites d'épissage consensus [47]. De plus 22 % des variants exoniques rapportés par HGMD peuvent altérer l'épissage [48], [49].

L'importance des variants splicéogéniques dans les maladies héréditaires s'explique par le fait que tout variant modifiant la séquence du pré-ARNm peut théoriquement altérer l'épissage (Figure 6). A l'instar de l'épissage alternatif, les remaniements du pré-ARNm lors de l'épissage peuvent être regroupés en différentes catégories.

Les variants situés dans les sites d'épissage consensus peuvent entrainer un saut d'exon par la perte de reconnaissance du site d'épissage. Ils peuvent aussi permettre l'utilisation d'un site cryptique non reconnu à l'état naturel avec la perte partielle de l'exon ou la rétention en partie de l'intron [50].

Les variants hors sites consensus peuvent aussi influencer l'utilisation de ces derniers par le biais des SREs. La perte des motifs *enhancers* (ISE/ESE) ou bien la création des motifs *silencers* (ISS/ESS) entraine des effets similaires aux variants dans les régions consensus sur l'épissage. La plupart des SREs étant situés au niveau exonique [25] explique la forte proportion de variants splicéogéniques au sein des exons.

De même que les variants situés dans les sites consensus, les variants situés au sein des points de branchement impactent également l'épissage avec la possibilité soit d'un saut d'exon, soit d'une rétention compète d'intron. Cependant les variants situés dans les points de branchement sont moins fréquemment rapportés du fait de la courte taille de motif et de la localisation du motif à distance du site accepteur [51].

La dernière catégorie des variants splicéogéniques est l'utilisation de nouveau site d'épissage. En effet, si l'altération des motifs d'épissage naturellement présents dans la séquence du pré-ARNm peut conduire à l'utilisation d'un nouveau site d'épissage, les variants peuvent directement engendrer de nouveaux sites d'épissage. Ainsi il a été rapporté des variants modifiant l'épissage même à distance des motifs d'épissage initiaux. A titre d'exemple, le variant c.759+26G>A du gène *PDHA1* (NM_000294) a été identifié comme entrainant la rétention de 45 nt de l'intron par utilisation d'un nouveau site donneur [52]. En théorie, la création de n'importe quel motif d'épissage, excepté les SREs *silencers*, peut entrainer l'utilisation d'un nouveau site d'épissage. Cependant, dans la grande majorité des cas l'utilisation d'un nouveau site d'épissage implique la création d'un site consensus d'épissage. Ce nouveau site d'épissage est qualifié soit de site *de novo* si le variant créé un site d'épissage inexistant à l'état naturel, soit de site cryptique s'il renforce un site préexistant [53].

Par ailleurs, un nouveau site d'épissage (ex : donneur) peut s'associer avec un site complémentaire (ex : accepteur) et ainsi former un nouvel exon [54]. Ces derniers, nommés pseudo-exons, sont principalement observés pour des variants situés à une grande distance (> 100 nt) des sites naturels d'épissage [55]. En effet pour le gène *DMD* (NM_004006), il a été observé la création d'un pseudo-exon jusqu'à plus de 30 000 nt du site d'épissage le plus proche [56].

La multiplicité des défauts d'épissage imputables à un variant splicéogénique rend d'autant plus complexe l'association entre variant et phénotype. En outre l'altération d'un même motif d'épissage peut conduire à plusieurs défauts d'épissage. Ces modifications d'épissage peuvent décaler le cadre de lecture en créant ainsi un PTC. Or la présence de PTC entraine la dégradation de l'ARNm par le NMD. Aussi, caractériser ces transcrits aberrants peut se révéler être un vrai défi.

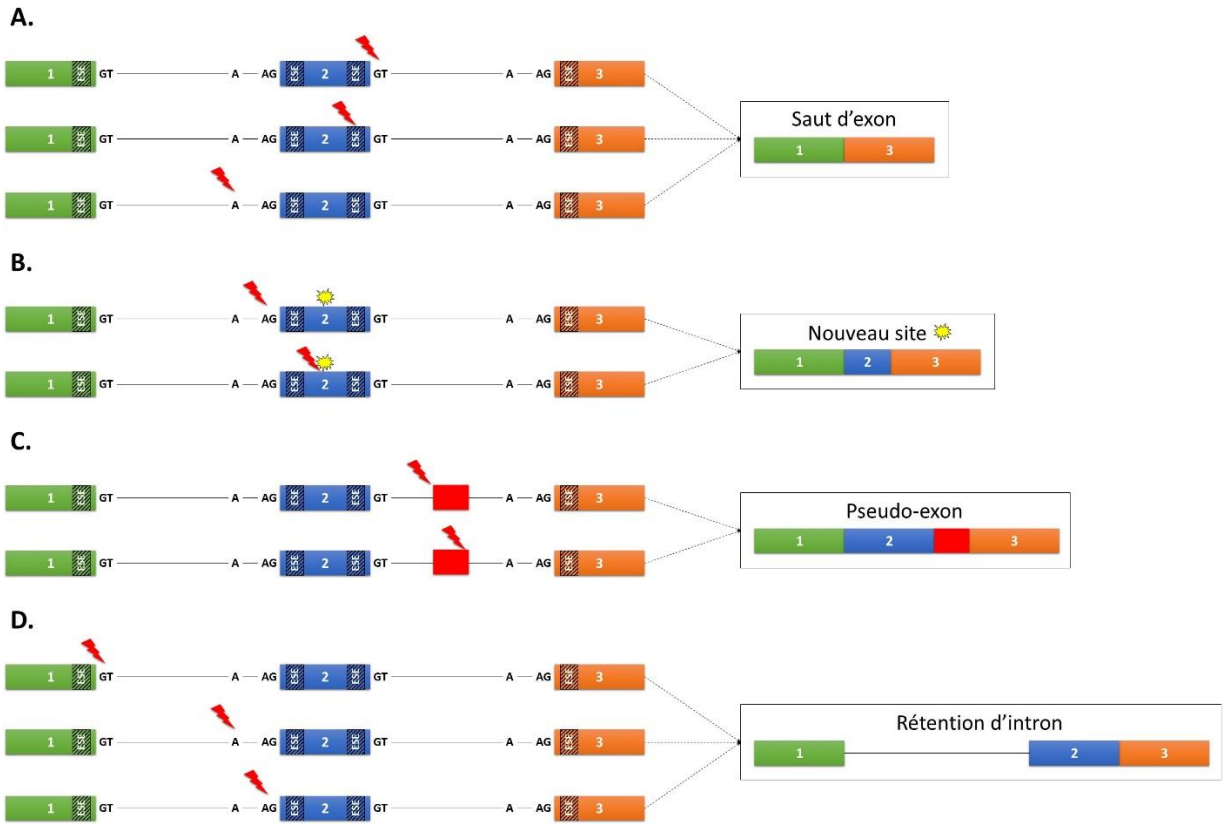


Figure 6 : Principales conséquences sur l'épissage de l'altération des motifs d'épissage par un variant. **A.** Illustration d'un saut d'exon résultant de l'altération de l'un des motifs d'épissage par un variant (éclair rouge) : donneur (GT) ou accepteur (AG), éléments régulateurs (ESE) ou point de branchement (A). **B.** Utilisation d'un nouveau site d'épissage soit par altération d'un site préexistant (ici un accepteur) et utilisation en relais d'un autre site (étoile jaune) ou soit activation d'un nouveau site par le variant. **C.** Apparition d'un nouvel exon (ou pseudo-exon) par l'activation de l'un de ses sites d'épissage. **D.** Rétention complète de l'intron par altération de l'un des motifs canoniques.

II. Tests fonctionnels dédiés aux défauts d'épissage

L'identification des défauts d'épissage représente un enjeu majeur en génétique moléculaire. Ainsi plusieurs stratégies ont émergé pour caractériser la séquence des ARNm. Nous pouvons distinguer deux approches différentes et complémentaires pour étudier les transcrits : celle utilisant les ARNm directement issus d'un échantillon biologique donné et celle permettant de reproduire artificiellement les transcrits d'intérêt. Du fait de l'essor du séquençage à haut débit depuis les années 2000, ces deux approches ont été adaptées pour être éligibles à des analyses à haut débit.

1. Les analyses *in vitro* à partir d'ARN naturel

a. Tests fonctionnels à bas débit

En parallèle de la découverte de l'épissage dans les années 1970, plusieurs techniques d'exploration de la séquence et de l'abondance des transcrits ont été proposées. Le *Northern blot* en est une des premières techniques. Adapté du *Southern blot* dédié à l'ADN, le *Northern blot* consiste en un séquençage par bande sur gel d'agarose [57]. L'analyse par electrophorèse permet de séparer les ARN en fonction de leur taille et la dernière base de la molécule est marquée. Nous pouvons également citer l'hybridation par points qui était une des premières méthodes pour quantifier l'expression des gènes [58]. Brièvement, cette méthode proposait d'utiliser des sondes ADN ou ARN couplées à des marqueurs radioactifs pour estimer à la fois le nombre de copies d'ADN génomique et le nombre de molécules ARNm, et ainsi pouvoir comparer l'expression des gènes entre eux.

Si ces techniques ont été délaissées au fil du temps, une autre technique apparue à la même époque, la RT-PCR pour *Reverse-Transcriptase Polymerase Chain Reaction*, est encore largement utilisée. Développée durant les années 1980, cette méthode reprend le principe de la *Polymerase Chain Reaction* (PCR) mise au point pour amplifier l'ADN dans le diagnostic de la drépanocytose [59]. Pour adapter ce protocole à l'ARN, il est nécessaire de procéder à une reverse transcription des ARN en ADN complémentaire (ADNc), rendue possible par la découverte des reverses transcriptases présentes chez les virus ARN [60].

La reverse transcriptase nécessite qu'une amorce s'hybride sur l'ARN pour pouvoir le reverse-transcrire en ADNc. Trois stratégies d'amorce sont utilisées avec chacune leurs forces et leurs faiblesses.

L'amorce peut se présenter comme étant spécifique d'un transcrit donné et dans ce cas seul ce transcrit d'intérêt sera rétro-transcrit. Cependant cette approche limite grandement la découverte de nouveaux transcrits.

Les amorces oligo-dT ont été mises au point afin de s'hybrider sur la partie poly-adénine (polyA) en 3' des ARNm matures. Ces amorces permettent la reverse transcription spécifique des ARNm en réduisant les biais de sélection des transcrits. Néanmoins, pour les longs ARNm, la reverse transcription peut ne pas s'effectuer efficacement pour la partie 5' de la molécule. Ce biais 5' est aussi dépendant du protocole utilisé et de la performance de la reverse transcriptase.

Le dernier type d'amorce consiste en l'utilisation de courts fragments hexamèriques de séquence aléatoire, aussi appelé *random hexamers*. Ces *random hexamers* ont la capacité de s'hybrider aléatoirement sur l'ensemble du matériel nucléotidique présent dans l'échantillon. Ainsi, le biais de sélection des transcrits et le biais de reverse transcription 5' sont évités. Toutefois, cette méthode de reverse transcription n'est pas spécifique des ARNm et la reverse-transcription n'étant plus en molécule unique, la structure complète des transcrits est perdue.

L'ADNc obtenu par ces différentes approches est ensuite utilisé pour être amplifié par PCR. Cette étape de PCR nécessite un couple d'amorce sens/anti-sens qui vont définir l'amplicon et sa spécificité. Brièvement, la PCR est constituée de plusieurs cycles d'amplification, chacun de ces cycles comprenant une étape de dénaturation, une étape d'hybridation des amorces et une étape d'élongation par l'ADN polymérase.

La RT-PCR présente l'avantage d'être facile à mettre en œuvre car elle ne nécessite qu'une trousse de réactifs et un thermocycleur. De plus, elle est capable de détecter des transcrits à partir d'une faible quantité d'ARN total (de moins d'1 µg à quelques ng). Aussi cette technique est devenu une méthode standard dans les laboratoires de biologie moléculaire tant en recherche qu'en diagnostic. De plus, cette méthodologie a été déclinée en de nombreuses versions selon les besoins des laboratoires. En effet une fois l'ADNc produit, la plupart des techniques de PCR développées pour l'ADN peuvent s'appliquer à l'ARN.

Parmi elles, nous pouvons citer les RT-PCR quantitatives, intégrant un fluorochrome à chaque production d'amplicon pour suivre en temps réel l'amplification de l'ADNc (revue par [61]). Les RT-PCR long-range par l'utilisation d'une ADN polymérase adaptée pour amplifier de grands fragments (~10-20 kb ou kilobases) [62]. Les RT-PCR multiplex intégrant plusieurs couples d'amorces pour amplifier simultanément plusieurs amplicons d'ADN [63]. Plus récemment il a aussi été développé des PCR dites digitales qui permettent également de quantifier précisément l'abondance du matériel nucléotidique présent dans l'échantillon. Brièvement, ce type de PCR consiste à diviser le milieu réactionnel en plusieurs micro-gouttes, chacune ayant un nombre aléatoire de séquence cible. Puis un signal lumineux est émis lors de l'amplification. La combinaison de l'ensemble de ces signaux permet d'estimer la quantité initiale de molécule d'intérêt (revue par [64]). En outre ces PCR digitales peuvent aussi s'appliquer pour l'études des ARN [65].

Le ou les amplicons obtenus à la suite des RT-PCR sont ensuite analysés au niveau de leur taille et de leur séquence. La taille des fragments est le plus souvent déterminée par simple migration sur gel d'agarose. Cependant une approche permettant d'être plus discriminante est l'électrophorèse capillaire. Initialement utilisée pour différencier des petites molécules, elle est maintenant employée pour différencier des amplicons dont la taille n'est séparée que de quelques nucléotides [66]. Ces derniers sont séquencés, avec ou sans une étape de purification sur gel des amplicons d'intérêt, selon une méthode décrite par Sanger en 1977 [67]. Celle-ci consiste à générer un ensemble de fragments de tailles aléatoires se terminant par un nucléotide marqué pour ensuite faire migrer ces fragments dont la taille et le dernier nucléotide permettent de retrouver la séquence initiale de l'amplicon. Ce procédé est maintenant automatisé dans des appareils d'électrophorèse capillaire comme l'ABI 3130xl d'Applied Biosystem®.

b. Tests fonctionnels à haut débit

Bien que les techniques de RT-PCR permettent de répondre à un certain nombre de questions, elles restent limitées de par le nombre de gènes et de transcrits pouvant être étudiés. Aussi à partir de la fin des années 1990, une nouvelle méthode d'analyse des transcrits a été proposée. Il s'agit des puces à ARN ou *microarrays* [68]. Leur principe est le suivant : les ARNm sont rétro-transcrits à partir de leur queue polyA, les ADNc obtenues sont marqués par des fluorochromes. Ces ADNc marqués sont déposés sur une plaque de verre présentant à sa surface plusieurs milliers de sondes réparties en cluster, chacun étant spécifique des transcrits d'un gène particulier. Puis la surface est lavée pour éliminer les ADNc non hybridés à la surface de la plaque. Celle-ci est ensuite exposée à la lumière pour exciter les fluorochromes et l'intensité de réémission de la lumière est mesurée pour chaque cluster. Après traitement du signal, l'expression des gènes peut être estimée. Cette technique s'est très largement répandue des années 90 à la fin des années 2000, pour établir des profils d'expression révélateurs d'un état physiopathologique particulier, notamment en cancérologie [69].

Cependant la limite majeure de cette technique est l'incapacité à caractériser dans son intégralité l'épissage des ARNm. En effet, seuls les transcrits pour lesquels des sondes ont été hybridées seront identifiés, ceci ne permettant pas la découverte de nouveaux transcrits. Cette limite technique a pu être levée en 2008, lorsque pour la première fois le séquençage à haut débit, jusqu'alors utilisé uniquement pour l'ADN, a été adapté au séquençage de l'ARN. Nommée RNA-seq, cette nouvelle technologie présente le double avantage de séquencer et de quantifier à haut débit l'ensemble du transcriptome.

Au début du séquençage à haut débit, trois grandes technologies étaient présentes (revue par [70]):

- L'amplification par pont d'Illumina®.
- Le séquençage sur billes de Roche®.
- La technologie solid proposée par Applied Biosystem®.

Cependant, la technologie Illumina® s'est largement plus démocratisée puisque près des ¾ des articles présents dans Pubmed traitant de *next-generation sequencing* ou NGS (terme anglophone pour séquençage à haut débit) utilisent la technologie Illumina® (Figure 7). Ainsi la plupart des protocoles RNA-seq reprennent cette technologie d'amplification par pont.

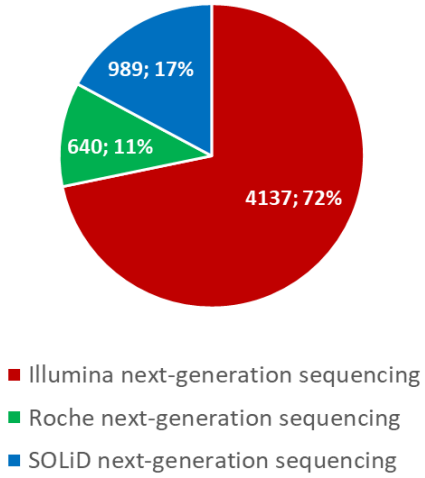


Figure 7 : Proportion d'articles présents dans Pubmed avec les mots-clés « *next-generation sequencing* » et le nom des principales technologies utilisées.

Typiquement, une expérimentation RNA-seq commence par l'enrichissement des ARN d'intérêt avec deux grandes approches : soit une purification des ARN polyA ou soit une déplétion des ARN ribosomaux. La purification des ARN polyA est peut-être la méthode la plus utilisée. La purification de ces ARN peut soit se faire par des billes magnétiques ou de cellulose recouverte de molécules oligo-dT ou bien par une reverse transcription *via* des amorces oligo-dT. A l'instar des RT-PCR utilisant les amorces oligo-dT, un biais de couverture en 5' de la molécule peut être également présent. La déplétion en ARN ribosomaux permet de conserver en plus des ARNm, d'autres types d'ARN pouvant relever d'un intérêt biologique (par exemple : micro-ARN, long ARN non codant). La plupart des protocoles utilisent des sondes permettant soit une sélection positive des ARN ribosomaux, qui sont ensuite dégradés par une RNase H, soit une sélection négative, les ARN d'intérêt étant reverse-transcrits et le reste dégradé.

A la suite de cet enrichissement en ARN d'intérêt, s'ensuivent deux étapes, la fragmentation et la reverse transcription pour les protocoles d'enrichissement n'impliquant pas déjà de reverse transcription. La fragmentation permet de réduire la taille des molécules car les technologies de séquençage à haut débit

dites « *short read* » ne permettent pas de séquencer des molécules de plus de 1 kb de long [71]. Cette fragmentions peut-être chimique, thermique et/ou enzymatique. L'étape de reverse transcription est spécifique ou non du brin initial de la molécule ARN. Une reverse transcription brin-spécifique permet de différencier les transcrits issus des gènes situés sur le brin sens ou anti-sens de l'ADN. Pour cela, la reverse transcriptase synthétise durant un premier cycle un seul brin d'ADNc auquel est ajouté un code barre, puis lors d'un second cycle le second brin d'ADNc est synthétisé et se voit doté d'un second code barre différent du premier. Cependant cette étape de reverse transcription n'est pas obligatoire car certains protocoles de RNA-seq propose de séquencer directement l'ARN. Cela est surtout utilisé pour les faibles quantités d'ARN ou bien en cas d'ARN dégradé [72].

Suite à ces étapes de purification, fragmentation et reverse transcription, les fragments d'ADNc sont ligués à de courtes séquences ADN dites adaptateurs. Ces derniers ont pour rôle de permettre l'hybridation des ADNc sur le support de séquençage de l'appareil. Au cours de cette étape une autre séquence ADN peut être liguée, appelée index. Celle-ci contrairement aux adaptateurs est de séquence variable. Les indexes permettent de pouvoir regrouper l'ensemble des échantillons dans une analyse RNA-seq, diminuant ainsi les coûts de séquençage.

Ainsi, ces fragments d'ADNc, appelés librairies, sont déposés sur le support de séquençage de l'appareil. Dans le cas de la technologie Illumina®, ce support de séquençage est une *flow cell*. Celle-ci présente à sa surface les séquences complémentaires aux adaptateurs précédemment ligués aux ADNc.

Une fois les librairies déposées sur la *flow cell*, l'appareil procède à une amplification par pont afin de générer des clusters dont les molécules sont identiques entre elles. Puis lors du séquençage, le brin complémentaire est synthétisé en utilisant des nucléotides marqués par un fluorochrome, de sorte qu'à chaque addition d'un nucléotide un signal lumineux correspondant au nucléotide est émis (Figure 8).

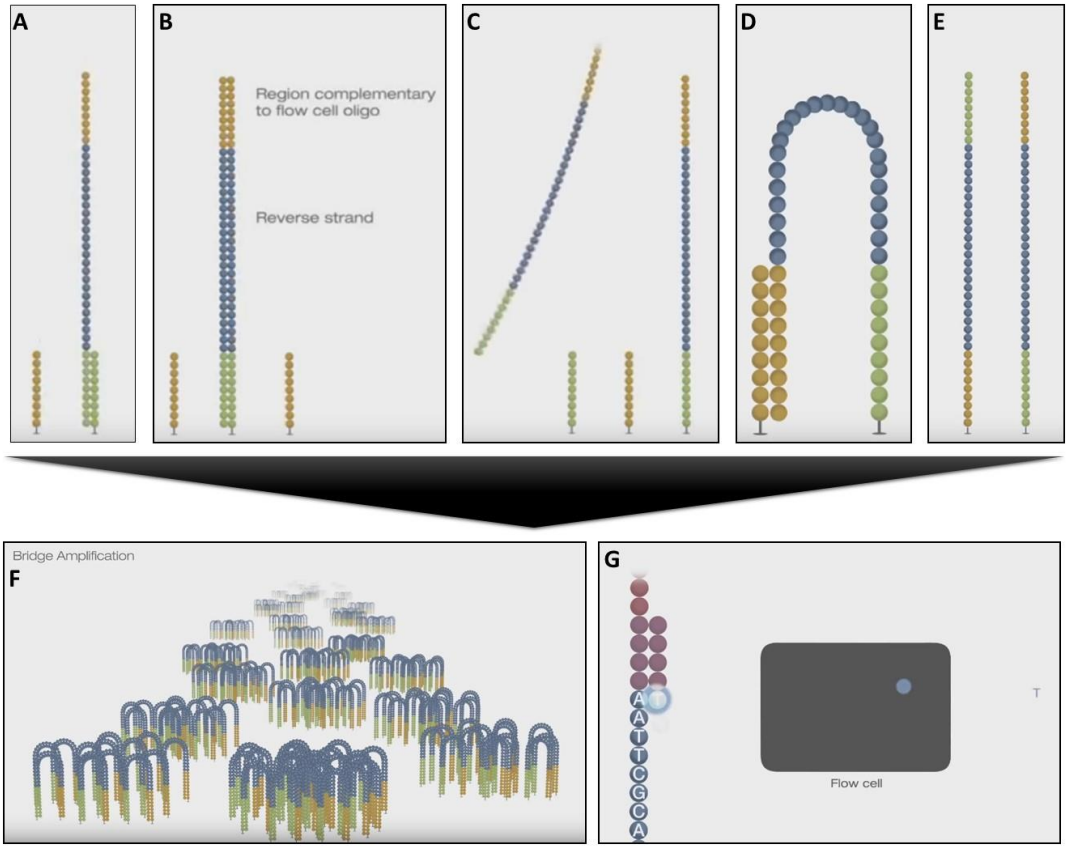


Figure 8 : Principe de l'amplification par pont proposé par Illumina (adaptée de Intro to Sequencing by Synthesis: Industry-leading Data Quality). **A.** La librairie s'hybride sur le support de séquençage grâce aux adaptateurs (vert et orange). **B.** Le brin complémentaire est synthétisé directement sur le support. **C.** La librairie est dissociée du brin néo-synthétisé. **D.** Formation d'un pont avec le second adaptateur. **E.** Deux brins sont produits à partir du pont formé. **F.** Répétition de l'amplification par pont pour former des clusters de même séquence. **G.** Séquençage par insertion de nucléotides marqués

La capacité de séquençage de ces appareils est définie par le nombre maximal de clusters possibles, étroitement lié à la taille de la *flow cell*, à la résolution de la caméra et à la longueur maximale de la molécule séquencée. Pour chaque cluster la séquence peut être lue une fois (séquençage *single-end*), séquence correspondant au brin sens, ou bien deux fois correspondent à la séquence du brin sens et anti-sens (séquençage *paired-end*). Le séquençage *paired-end*, dont les deux *reads*, aussi appelées *reads*, permet d'améliorer la qualité du séquençage en corrigeant de possibles erreurs de lecture apparues sur l'un des brins. Ainsi, à titre d'exemple, l'appareil NextSeq 500 d'Illumina® peut lire jusqu'à 400 millions de clusters, soit 400 millions de *reads* en *single-end* et 800 millions de *reads* en *paired-end*. En parallèle, cet appareil peut séquencer des molécules jusqu'à 150 nt de long. Ainsi la capacité de séquençage en nucléotide est le produit de ces deux caractéristiques, soit 150 nt x 800 millions de *reads* ce qui donne 120 $x10^9$ nt ou 120 Gb. Cette capacité de séquençage correspond à environ 40 fois la taille du génome humain (~3.2 Gb). Or la dernière plateforme de séquençage proposée, le NovaSeq 6000, peut lire jusqu'à 10 milliards de clusters (20 milliards de *reads* en *paired-end*) pour une longueur de molécule séquencée

jusqu'à 150 nt. Ainsi la capacité de séquençage maximale actuellement disponible est de 3 000 Gb soit près de 1 000 fois la taille du génome humain.

Suite à l'essor du RNA-seq au cours des dix dernières années, de nombreuses versions du RNA-seq ont été proposées, notamment le single-cell RNA-seq (scRNA-seq) et le RNA-seq ciblé. Le scRNA-seq est particulièrement intéressant pour identifier un profil d'expression d'une cellule ou d'un sous-type cellulaire (revue par [73]).

Le RNA-seq ciblé regroupe l'ensemble des techniques permettant d'enrichir certains transcrits d'intérêt. Cette technique, très peu adaptée pour définir un profil d'expression, est particulièrement intéressante pour étudier les gènes peu exprimés ou définir un patron d'épissage alternatif d'un panel de gènes. En effet l'augmentation de la couverture de séquençage, c'est-à-dire le nombre de *reads* par position, permet de détecter des transcrits alternatifs exprimés à bas bruit dans la cellule. Deux méthodes ont été mises en place pour cibler les transcrits d'intérêt, soit un enrichissement par amplification, soit un enrichissement par capture. La méthode par amplification est la plus simple. Elle consiste à réaliser une PCR spécifique du panel de transcrits avant le séquençage. Les PCR multiplex et long-range sont particulièrement adaptées pour ce type de méthode [74]. Cependant cette approche génère de nombreux biais, que ce soit lors de la fixation des amorces ou bien lors de l'amplification à proprement parler. L'approche par capture permet de réduire ces biais bien qu'elle soit plus coûteuse [75]. La capture des transcrits implique l'utilisation de sondes, le plus souvent couplées à de la biotine, venant s'hybrider sur les librairies d'intérêt (Figure 9) [76]. Le complexe sondes-librairies est ensuite positivement sélectionné par des billes magnétiques recouverte de stréptavidine, celle-ci étant affine pour la biotine. Après élution, les librairies capturées sont utilisées pour le séquençage. L'enrichissement, plus le maintien de la diversité des transcrits, font du RNA-seq ciblé par capture, une méthode de choix pour explorer les différents ARNm issus de l'épissage alternatif [77].

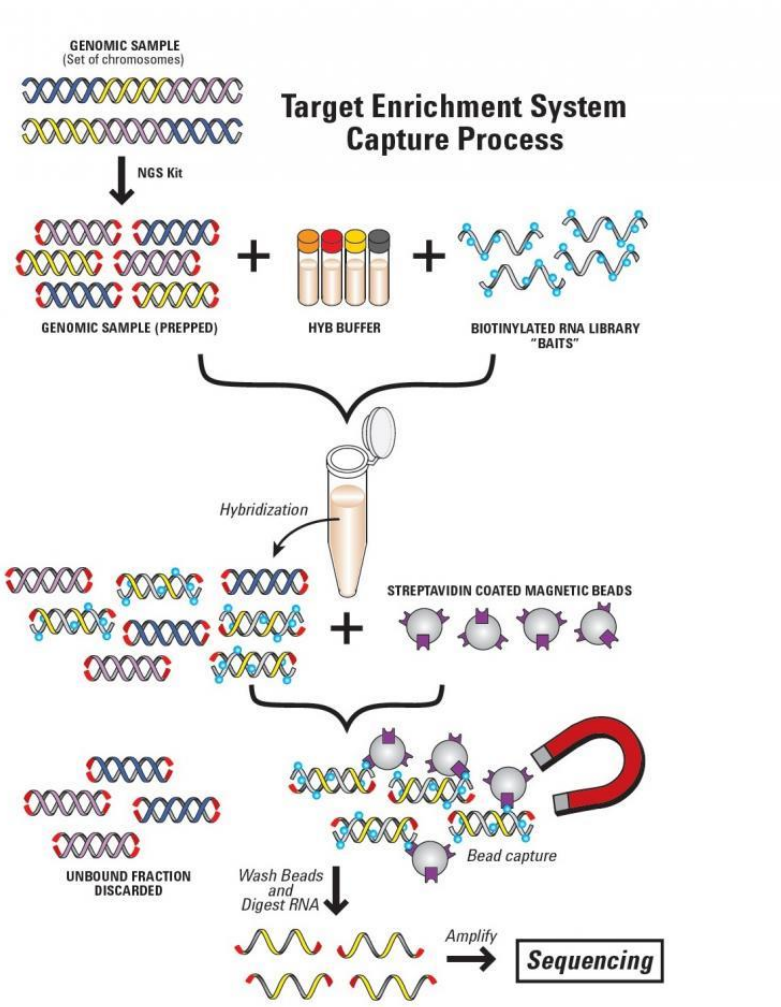


Figure 9 : Illustration du principe de capture des librairies avec le protocole SureSelect XT d'Agilent®. Les librairies sont associées à des sondes biotinylées. Après hybridation, des billes magnétiques recouvertes de stréptavidine sont intégrées dans le milieu réactionnel. Le complexe librairie-sonde-bille est positivement sélectionné par un aimant. Après lavage, les librairies capturées sont utilisées pour le séquençage.

Cependant le RNA-seq tel qu'il a été conçu contient une limitation technique. En effet, une étape de fragmentation est nécessaire pour pouvoir séquencer les ARNm. Aussi la détection des évènements d'épissage se fait toujours de manière isolée. Il en résulte que l'assemblage complet d'un ARNm par l'épissage, appelé également phase du transcrit, est méconnu.

En réponse à cette limitation, il est apparu au début des années 2010 une nouvelle génération de séquençage à haut débit. Cette nouvelle génération, nommée *third generation sequencing* par opposition au *next-generation sequencing*, ou séquençage « *long-read* », propose de séquencer des longs fragments de 1 kb à 100 kb. Parmi les technologies proposées, nous pouvons citer celles de Oxford nanopore® et Pacific Bioscience ou PacBio®.

Oxford nanopore® utilise le principe de *ion torrent* ou flux d'ion [78]. Ce dernier consiste à faire traverser le fragment d'ADN par un pore situé sur une membrane séparant un gradient de concentration en ion. Lors que la molécule traverse ce pore, par encombrement stérique, chaque nucléotide va plus ou moins perturber le courant ionique (Figure 10). Ces fluctuations de courant sont enregistrées puis converties en séquence ADN correspondante.

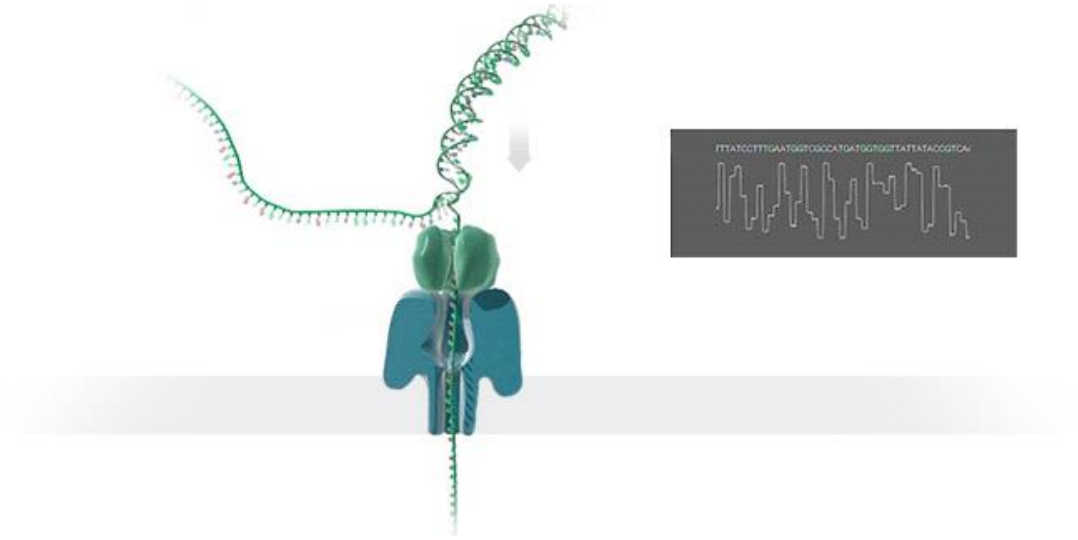


Figure 10 : Technologie de séquençage *long read* utilisée par Oxford nanopore®. Le brin d'ADN passe par un pore situé sur une membrane avec de part et d'autre une différence de tension. Chaque nucléotide freinant plus ou moins le passage du courant, l'appareil mesure la différence de tension électrique pour identifier le nucléotide.

La technologie PacBio® repose sur le séquençage SMRT (*Single-molecule real-time*). Le séquençage SMRT utilise une ADN polymérase fixée au fond d'un puit (Figure 11) [79]. Cette enzyme accroche le fragment d'ADN pour en synthétiser le brin complémentaire. Or les nucléotides utilisés pour cette synthèse sont marqués par un fluorochrome de sorte qu'à chaque addition d'un nucléotide, un signal lumineux est émis. Le séquençage à haut débit de deuxième génération repose sur un signal émis par un cluster et donc de plusieurs molécules.

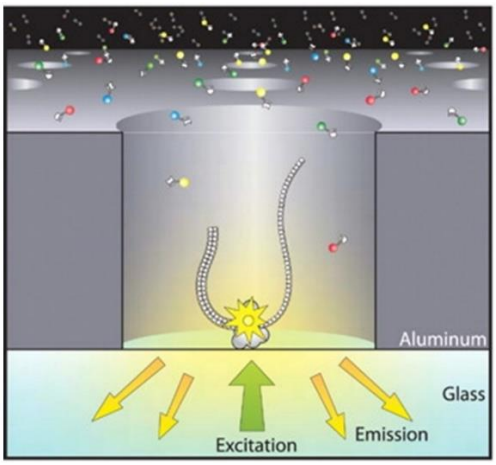


Figure 11 : Principe de la technologie utilisée par Pacific Bioscience® pour le séquençage *long read*. Le support de séquençage est constitué de puits avec pour chacun une ADN polymérase. Cette polymérase capture une librairie pour en synthétiser le brin complémentaire avec des nucléotides marqués. Après excitation de ces nucléotides le signal émis est détecté.

Ici, pour les deux technologies, le signal émis pendant le séquençage ne provient que d'une seule molécule. Aussi ce signal est très proche du bruit de fond et de nombreuses erreurs peuvent être générées. Si la technologie *ion torrent* propose peu de solution technique pour corriger ces erreurs. Le séquençage SMRT circularise les librairies avant que ces dernières ne soient déposées sur le support de séquençage. Ainsi la molécule est séquencée en boucle pour former un *read* chimérique contenant autant de fois la séquence de la librairie que le nombre de passage qu'a réalisé l'ADN polymérase. Puis informatiquement ces séquences sont auto-alignées afin de corriger les erreurs de séquençage. Ces technologies ont été adaptées pour le séquençage des ADNc et pour le séquençage ARN direct pour *ion torrent*. Ainsi le séquençage *long-read* offre de nouvelles possibilités pour identifier des mécanismes complexes d'épissage [80], [81]. D'autant plus que de nouvelles stratégies d'enrichissement émergent pour améliorer la détection des panels d'épissage des transcrits d'intérêt [82], [83]. Cependant ces dernières sont uniquement basées sur un enrichissement par PCR, avec un risque important de biais de sélection.

2. Les analyses *in vitro* à partir d'ARN artificiel

L'ARN directement issu des échantillons biologiques est privilégié pour les analyses d'épissage. Cependant, un certain nombre de limitations techniques peuvent se poser. Tout d'abord l'ARNm est une molécule instable, se dégradant en quelques heures à température ambiante. De plus dans le cadre de la routine clinique l'obtention de tissus frais peut se révéler être un vrai challenge. En outre, les quantités d'ARN obtenus à partir des tissus sont souvent faibles au regard de celles obtenues par culture cellulaire. L'identification de transcrits aberrants hors phase, associés à un variant splicéogénique, est également entravée par la dégradation de ces derniers *via* le NMD. Dans le cas de variants splicéogéniques

hétérozygotes, la distinction entre les transcrits produits par les deux allèles est complexe. C'est notamment le cas lorsque le défaut d'épissage lié au variant est l'amplification d'un épissage alternatif naturellement présent dans la cellule.

Pour répondre à ces limitations techniques, il a été proposé, concernant la dégradation de l'ARN, de tester son intégrité par un critère d'évaluation appelé *RNA Integrity Number* (RIN). Le principe de ce critère est de calculer le ratio entre les ARN ribosomaux 18S/28S [84]. Normalement ce ratio est supérieur ou égal à 2. Historiquement, la quantité d'ARN était mesurée en faisant migrer l'échantillon sur gel d'agarose. Cependant l'estimation de ces quantités restait relativement arbitraire. Aussi l'essor du système micro-fluidique a permis de mettre en place un système de mesure plus fiable du RIN, le premier système micro-fluidique étant le bioanalyser 2100 d'Agilent® [85].

Pour faciliter le prélèvement de tissu dans le cadre de la routine clinique, des systèmes de conservation de l'ARN ont été mis en place. Un des plus utilisés étant le tube PAXgene™ permettant notamment de conserver l'ARN de tissu sanguin pendant 3 jours à température ambiante ou 5 jours à 4°C [86]. Cependant pour les tissus solides, peu de solutions existent pour conserver l'ARN. De plus, la fixation des tissus entraîne le plus souvent la dégradation des ARN [87]. Pour réduire l'impact du NMD, l'utilisation de différents inhibiteurs a été suggérée. La plupart sont des antibiotiques tels que anisomycine, cycloheximide, emetine, pactamycine, et puromycine, agissant notamment en inhibant la traduction [88], [89]. Cependant l'utilisation de ces inhibiteurs du NMD nécessite la mise en place d'une culture cellulaire, ce qui n'est pas toujours possible selon le type d'échantillon biologique étudié. En outre il n'y a pas de protocole standard pour les utiliser, sachant que leur action est à la fois dépendante du temps d'exposition et de leur concentration [90]. Ces inhibiteurs n'étant pas spécifiques du NMD, ils peuvent modifier l'expression d'autres gènes et donc influencer les résultats finaux par modification du profil d'expression.

La difficulté de répondre à l'ensemble de ces contraintes a incité les laboratoires à développer des systèmes permettant de reproduire artificiellement les transcrits d'intérêt.

a. Tests fonctionnels à bas débit

Au cours de la seconde moitié du 20ème siècle, les technologies de clonage moléculaire ont connu une extension et une démocratisation particulièrement importante [91]. Ainsi l'idée de créer un plasmide (ADN circulaire) codant pour un transcrit d'intérêt a émergé dès les années 1980. Appelé test minigène, ces constructions reprennent la construction standard d'un plasmide :

- Promoteur : pour initier la transcription
- Gène de résistance : pour sélectionner positivement les cellules transfectées

- Sites de restriction : séquence ADN reconnue par les enzymes de digestion pour ajouter un insert, une séquence ADN personalisée.

Le plus souvent, l'insert correspond à l'exon affecté par le variant, plus les séquences introniques bordantes. Parmi la diversité des systèmes minigènes, nous pouvons citer le plasmide pCAS2 [92]. Cette construction contient deux exons du gène *SERPING1/C1NH*, avec dans l'intron central les sites de restriction *BamHI* et *MluI*. Ces sites de restriction sont utilisés pour cloner l'exon d'intérêt, plus les séquences introniques bordantes de +/- 150 nt (Figure 12). Une fois l'insert cloné dans le plasmide, ce dernier est transfecté dans des cellules HeLa. L'ARN est ensuite extrait, puis par RT-PCR spécifique des transcrits produit par le plasmide, le défaut d'épissage peut être révélé.

Les tests minigènes présentent l'avantage de produire une grande quantité d'ARN et de bonne qualité. Pour certains minigènes l'insert est issu d'un fragment d'ADN génomique du patient. Cependant il est possible de recréer artificiellement le variant par mutagénèse dirigée afin de s'affranchir de tout échantillon biologique provenant de l'individu. De plus il est possible de développer des constructions insensibles au NMD, par exemple en supprimant le codon d'initiation ATG du gène.

Cependant une des principales questions qui peut se poser avec les systèmes minigènes est la concordance qualitative et quantitative des transcrits entre le système minigène et le tissu d'intérêt. En effet, seule une partie du gène est cloné dans le plasmide et les lignées cellulaires transfectées sont le plus souvent non spécifiques des modèles cellulaires du tissu d'intérêt (ex : HeLa, HEK293, …). Ainsi plusieurs études ont confronté les résultats fournis par les systèmes minigènes et ceux observés chez l'individu. Dans certains cas, des discordances ont pu être rapportées entre ces deux approches [93], [94]. Cependant à grande échelle, il n'est pas observé de différence majeure entre les minigènes et les ARN naturels [95], [96]. Ainsi les minigènes représentent un moyen pertinent pour étudier l'épissage, lorsque l'ARN naturel n'est pas accessible.

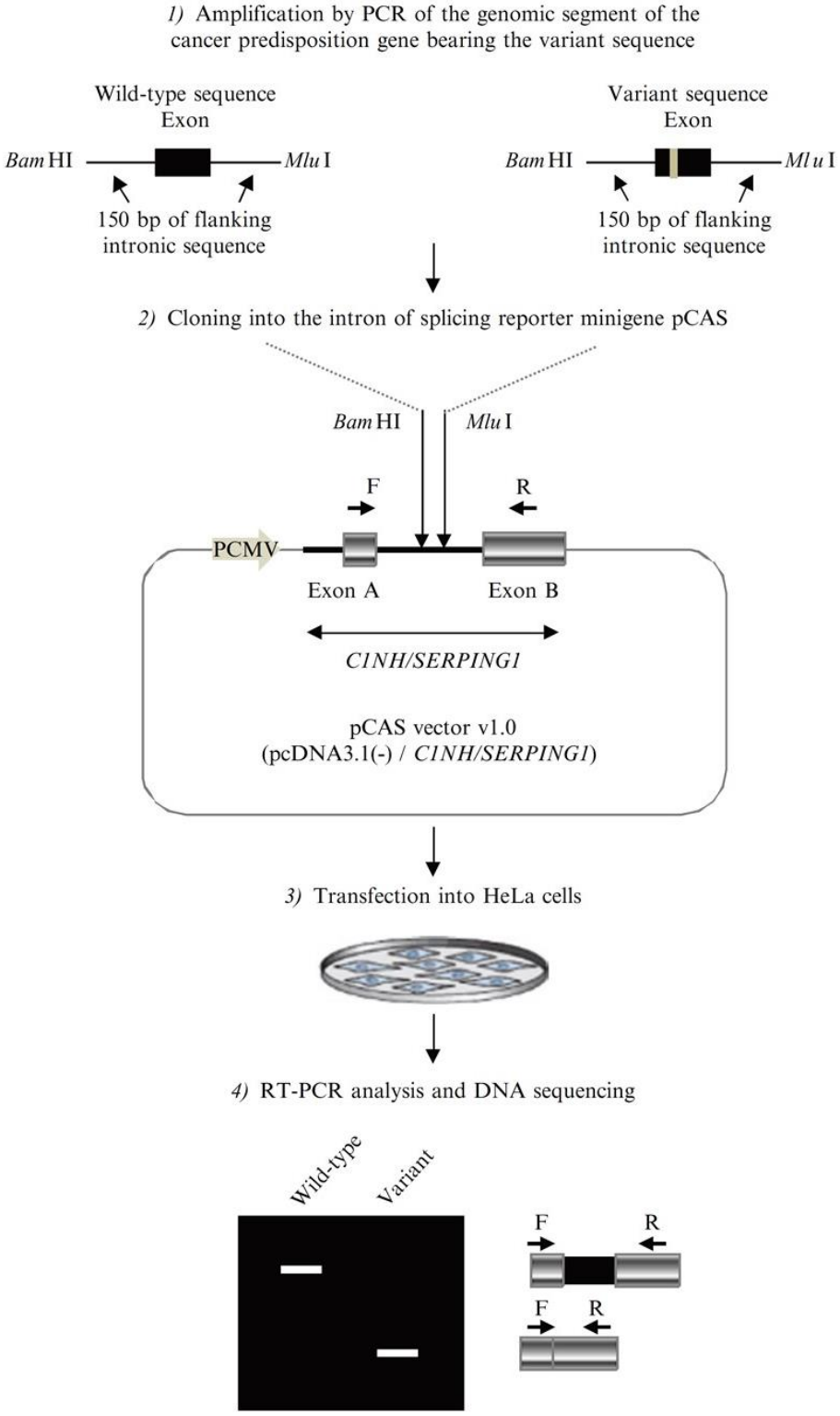


Figure 12 : Exemple d'un test minigène avec le plasmide pCAS1 (d'après [97]). **(1)** la séquence exonique sauvage et mutée sont amplifiées par PCR à partir d'ADN de patient, en utilisant des amorces avec les sites de restriction BamHI et MluI. **(2)** Les amplicons sont clonés dans le plasmide pCAS1 contenant un minigène composé de deux exons (ici, nommé A et B) du gène CINH/SERPING1. **(3)** et **(4)** Après transfection de la construction pCAS dans des cellules HeLa, l'ARN total est extrait puis les transcrits sont analyses par RT-PCR, en utilisant les amorces sens (F) et anti-sens (R) complémentaires des exons A et B.

b. Tests fonctionnels à haut débit

La génération des plasmides pour les tests minigènes s'avère être longue (2 à 3 semaines) avec un risque d'échec. Aussi rapidement il a été proposé de ne pas réaliser les tests minigènes de façon unitaire mais plutôt d'en générer un très grand nombre et de les tester en parallèle grâce aux récentes techniques de RNA-seq. Ainsi, au début des années 2010, les *massively parallel reporter assays* (MPRAs) ont émergé en combinant les test minigènes et le RNA-seq [98]. Leur principe est le suivant, sur une séquence d'intérêt plusieurs variants sont générés. Il peut y avoir un ou plusieurs variants par séquence, voire un motif aléatoire, avec pour chaque séquence l'ajout d'une séquence barre-code identifiante. L'ensemble de ces séquences est ensuite cloné dans des plasmides. Après transfection et culture cellulaire, les ARN produits par ces plasmides sont isolés puis séquencés par RNA-seq (Figure 13). Ces approches de MPRAs ont été notamment largement utilisées pour décrire l'environnement en SREs autour des sites d'épissage [24], [25]. L'étude de ces motifs impliquait la génération de motif aléatoire autour d'un système exon/intron connu et bien défini, usuellement 3 exons séparés par 2 introns. Pour chaque construction, le niveau d'inclusion de l'exon était mesuré. Ainsi chaque niveau d'inclusion d'exon est associé à une séquence particulière. En conséquence, les motifs ayant le plus d'impact sur l'épissage ont pu être identifiés.

Pour compléter ces technologies, de nouveaux protocoles ont récemment été développés pour mesurer l'impact de variant sur l'épissage à haut débit. La technologie MaPSy pour *massively parallel splicing assay*, consiste à recréer artificiellement l'ADN des patients, comprenant le système 3 exons affecté par un variant. Puis une séquence commune est ajoutée, celle-ci permettant de cloner l'ensemble de ces ADN dans le même système plasmidique. Plusieurs milliers de variants sur plusieurs systèmes exons/introns peuvent donc être étudiés simultanément. L'équipe de William Fairbrother, ayant développé cette technique, a pu étudier près de 5 000 variants présents dans la base de données HGMD [99]. En parallèle de cette technique, l'outil Vex-seq, pour *variant exon sequencing*, propose également de mesurer l'impact sur l'épissage par une approche similaire à MaPSy [100]. Les auteurs de cet outil ont pu étudier plus de 2 000 variants sur plus de 100 exons différents. Plus récemment une autre étude a proposé en janvier 2019 la méthodologie MFASS ou *multiplexed functional assay of splicing using Sort-seq*. Cette approche a d'original que le défaut d'épissage n'est pas mesuré par RNA-seq mais par fluorescence [101]. La séquence exonique affectée par le variant est insérée entre deux exons, codant pour les parties N-terminale et C-terminale de la GFP (*green fluorescent protein*). Si l'exon est inclus dans le transcrit alors la GFP est non fonctionnelle. Alors que le saut de l'exon permet la production d'une GFP fonctionnelle dont la fluorescence est mesurée par cytométrie en flux. La technologie MFASS a ainsi permis d'étudier 27 733 variants présents dans plus de 2 000 exons. Les variants utilisés étaient issus de la base de données ExAC (Exome Aggregation Consortium), base de données regroupant les variants exoniques et juxta-exoniques chez plus de 60 000 personnes [102].

Si ces technologies de tests fonctionnels à haut débit offrent la possibilité d'étudier un grand nombre de variants au niveau de l'épissage, elles sont limitées dans les défauts d'épissage mis en évidence. En effet, ces approches sont surtout conçues pour détecter des sauts d'exons (MaPSy et MFASS) et/ou l'utilisation d'un nouveau site d'épissage à proximité des sites naturels (Vex-seq). Les sauts d'exons multiples, les pseudo-exons et les rétentions d'intron ne peuvent être identifiés. Aussi ces techniques sont peu utiles pour caractériser finement le défaut d'épissage d'un ensemble de variants mais sont particulièrement intéressantes pour objectiver la proportion de variants splicéogéniques au sein d'une population donnée. En effet, MaPSy a ainsi démontré qu'environ 10 % des variants HGMD étudiés modifiaient l'épissage. MFASS a aussi dévoilé que 3.84 % des variants ExAC impactait l'épissage et parmi ces derniers seulement 17 % étaient situés sur un site canonique.

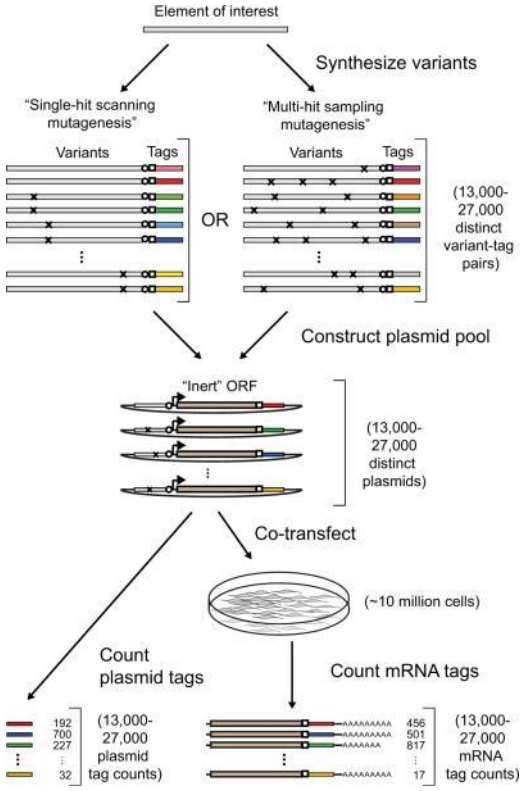


Figure 13 : Illustration du principe général des *massively parallel reporter assays* (MPRAs) [98]. Les différents fragments d'ADN contenant les variants sont synthétisés et couplés à des tags spécifiques. Les tags et les séquences sont séparés par deux sites communs de restriction (cercle/carré). Après amplification par PCR, les produits de PCR sont clonés dans un plasmide. Puis un promoteur est inséré en amont des séquences par double digestion suivi d'une liguation. Les plasmides sont co-transfectés dans une population cellulaire. L'expression relative est déduite par séquençage puis comptage des séquences à partir des ARNm.

III. Les outils bioinformatiques et biostatistiques dédiés au RNA-seq

L'essor des technologies de séquençage à haut débit a très vite posé la problématique de l'analyse des données fournies par séquençage. Pour se figurer la volumétrie des données générées, nous pouvons prendre pour exemple les données fournies par un séquenceur NextSeq 500, soit 800 millions de *reads* pour un volume final de 120 Gb. Pour écrire ces données sur papier, il faudrait 288 tonnes de feuille A4, soit le poids de 7 camions semi-remorques. Ce chiffre est d'autant plus vertigineux que la masse moyenne d'ADN déposée sur le support de séquençage est d'environ 200 ng, soit un rapport de poids entre l'ADN et le papier de l'ordre de 10^{15}. De plus une analyse manuelle des *reads* du séquenceur, à raison de 10 secondes par *read*, prendrait environ 250 ans. Aussi l'automatisation des analyses de ces données est devenue un enjeu majeur et a encouragé le développement de nombreux outils bioinformatiques et biostatistiques.

1. Les outils bioinformatiques

Il est difficile de définir un ou plusieurs outils bioinformatiques optimaux étant donné la diversité des applications du RNA-seq et des analyses possibles. Ainsi dans ce chapitre nous nous focaliserons sur les principales étapes d'analyses des données issues du RNA-seq, à savoir, l'alignement, l'annotation et le comptage. Cette dernière étape de comptage sert de point d'entrée pour la plupart des analyses biostatistiques.

a. Format des principaux fichiers utilisés en bioinformatique

Devant la volumétrie des données générées par le séquençage à haut débit et pour permettre l'automatisation des analyses, il a été proposé d'homogénéiser les formats des données en fonction de leur nature. Pour représenter les séquences en acides nucléiques, les formats Fasta et FastQ sont utilisés. Le format Fasta, développé par David Lipman en 1988 [103], est un fichier texte d'une seule colonne comprenant le nom de la séquence et le code de la séquence suivant la nomenclature usuelle IUPAC (*International Union of Pure and Applied Chemistry*) [104]. Le nom des séquences est indiqué par la présence du caractère « > » en première position, le reste des lignes étant considérées par défaut comme le contenu de la séquence. Plus utilisé pour représenter les données de séquençage, le format FastQ contient 3 types d'informations, le nom de la séquence et son contenu ainsi qu'un code de qualité de

séquençage pour chaque base. Ce code de qualité est basé sur le score PHRED, ce dernier étant la transformation logarithmique de la probabilité d'erreur de séquençage [105] :

$$Q_{PHRED} = -10 \times \log_{10}(P_{erreur})$$

A titre d'exemple une probabilité d'erreur de 5 % correspond à un score PHRED de 13. La valeur du score PHRED est représentée à la suite de la séquence concernée en code ASCII (*American Standard Code for Information Interchange*) (https://www.ascii-code.com/), le numéro du code ASCII étant défini par le score PHRED. Pour les plateformes de type Illumina, le numéro de caractère du code ASCII est défini ainsi : PHRED + 64 [106]. Ainsi un score PHRED à 13 est représenté par le caractère ASCII numéro 77 auquel correspond le code « M » (Figure 14). Cette stratégie a pour but de réduire le nombre de caractères et d'éviter des caractères vides ou non imprimables.

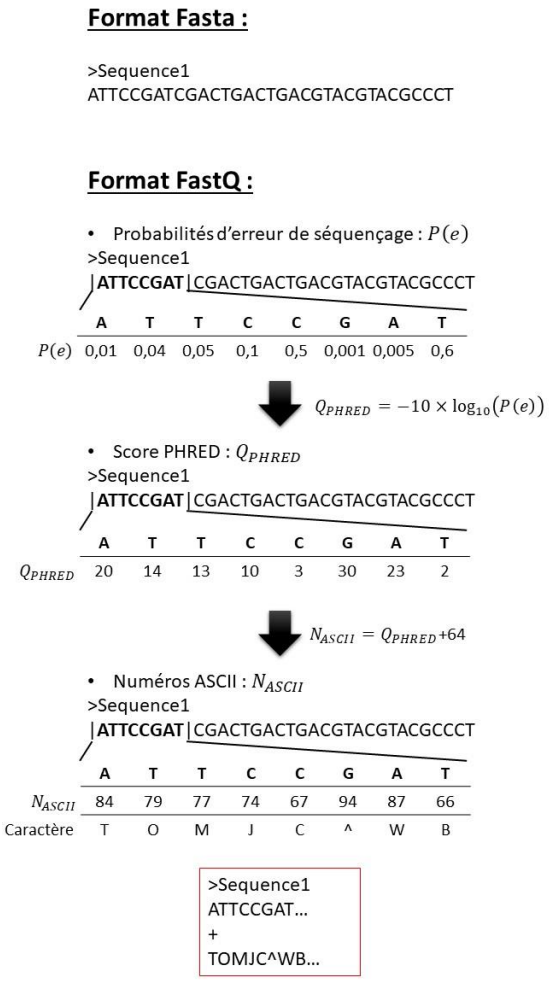


Figure 14 : Principe des fichiers Fasta et FastQ. Les fichiers Fasta ne comprennent que la séquence ADN ; Les fichiers FastQ comprennent la séquence plus la probabilité d'erreur de séquençage convertie en score PHRED puis en code ASCII (*American Standard Code for Information Interchange*).

Si les fichiers Fasta et FastQ renseignent sur le contenu de la séquence et sur la qualité de séquençage, ils ne contiennent pas d'information sur la position de la séquence au sein du génome. Pour cela, il est utilisé les fichiers dit d'alignement, qui en plus de la séquence et du score de qualité, intègrent la position de la séquence. Les standards actuellement utilisés sont les fichiers SAM/BAM pour *Sequence Alignment Map* et *Binary Alignment Map*. Les fichiers SAM sont des fichiers texte dont les colonnes sont séparées par des tabulations. L'en-tête des fichiers est identifiée par le symbole « @ ». Pour chaque lecture ou *read*, 11 champs doivent être renseignés incluant l'identifiant du *read*, la première position de la séquence, la taille de celle-ci, le contenu de la séquence et les scores PHRED [107]. Les fichiers BAM sont une version compressée des fichiers SAM, ayant les mêmes informations que ces derniers.

Pour représenter les informations concernant la structure du génome ou du transcriptome, il est utilisé les fichiers BED, ou *Browser Extensible Data*, et GTF/GFF (*General Transfer Format*/*General Feature Format*). Les fichiers BED ont été développés au début des années 2000 pour permettre la visualisation des structures génomiques et transcriptomiques par l'UCSC (*University of California Santa Cruz*) *Genome Browser* [108]. Le format BED comprend un fichier texte de 3 à 12 colonnes séparées par des tabulations (voir exemple en Figure 15). Les formats GTF/GFF sont eux, plus détaillés dans les informations fournies. En effet, en plus du nom et des coordonnées de la structure, ils fournissent aussi le détail de son rôle fonctionnel le cas échéant, les correspondances avec d'autres bases de données, *etc.* Etant donné ce plus grand nombre d'information, les fichiers GTF/GFF comprennent plusieurs lignes pour une même structure. La différence entre les fichiers GTF et GFF correspond aux numéros de version de ces fichiers. Actuellement les GFF sont sous la version 3, alors que les GTF sont sous la version 2 des GFF.

En parallèle de ces fichiers standards, il a été établi en 2011 un format dédié aux variants génétiques, nommé VCF pour *Variant Call Format* [109]. Le format VCF est un fichier texte d'au moins 8 colonnes séparées par des tabulations. Ces colonnes contiennent la position génomique du premier nucléotide muté, le nom du variant et la séquence sauvage et mutée (Figure 16). Ces séquences sont toujours rapportées par rapport au brin sens.

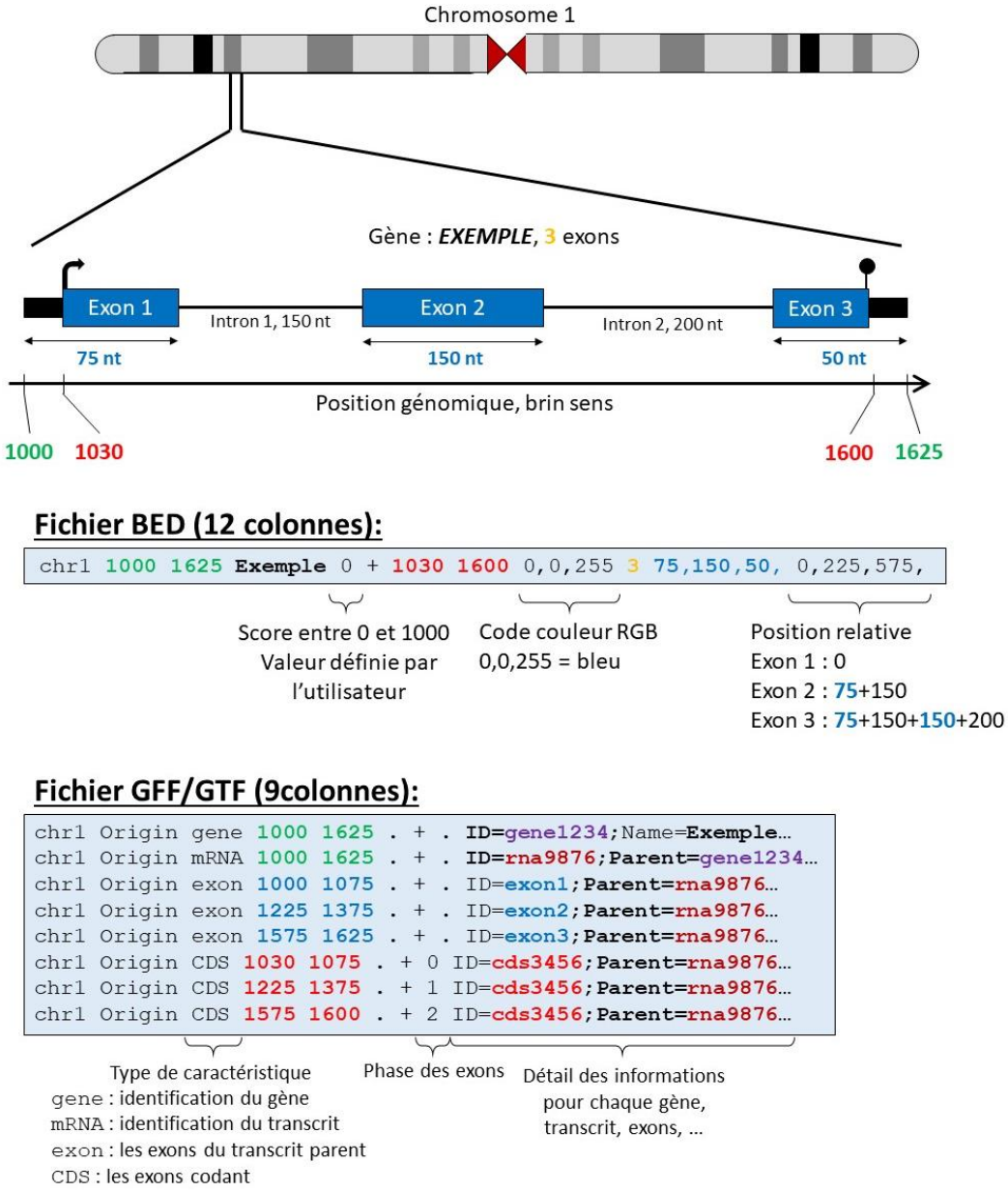


Figure 15 : Illustration des informations contenues dans un format BED et un format GTF/GFF, avec un exemple de transcrit ayant 3 exons.

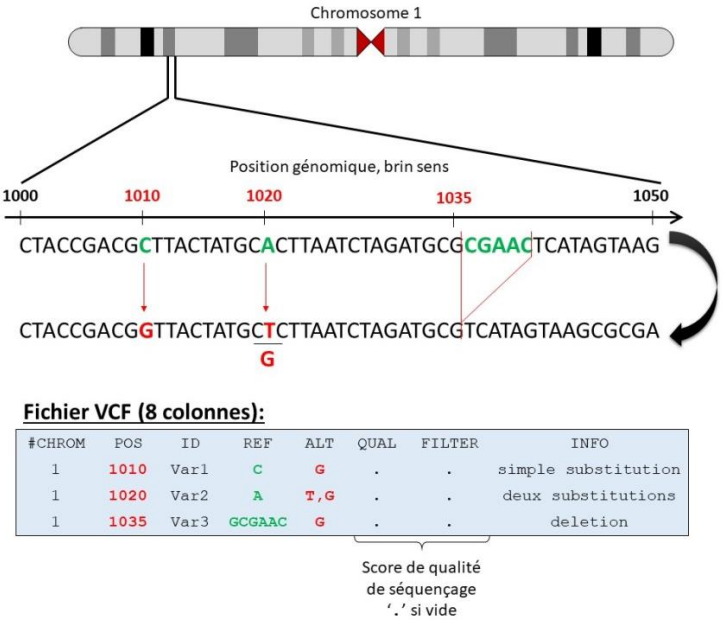


Figure 16 : Présentation des informations contenues dans un fichier VCF pour décrire les variants génétiques.

b. Alignement des données RNA-seq

L'alignement est la première étape d'analyse des données brutes et est indispensable pour l'identification et le comptage des transcrits. En effet durant cette étape les *reads* du séquenceur sont assemblées pour définir les transcrits initiaux présents dans l'échantillon. L'assemblage de ces *reads* peut se faire avec ou sans un fichier de référence. Les fichiers de référence sont des fichiers Fasta contenant l'ensemble des séquences génomiques ou transcriptomiques de l'espèce considérée.

Cependant ces fichiers ne sont pas toujours disponibles ou complets, c'est pourquoi des outils pour l'assemblage *de novo* des transcrits ont été proposés. Par exemple l'outil Trinity [110] permet l'assemblage *de novo* en regroupant les *reads* en contigs. Ces contigs sont des groupes de *reads* chevauchants entre eux. Puis les contigs sont eux-mêmes regroupés en cluster puis organisés entre eux, ici en utilisant l'approche de Bruijn [111]. Ainsi les différentes isoformes peuvent être reconstruites (Figure 17). A noter que ces reconstructions sont d'autant plus fiables si les *reads* initialement utilisés sont longs. C'est pourquoi le séquençage *long-read* est particulièrement adapté pour l'assemblage *de novo* [112]. A ce titre PacBio fourni, en plus du séquençage SMRT, un pipeline nommé Iso-Seq basé sur l'assemblage *de novo* pour étudier les isoformes.

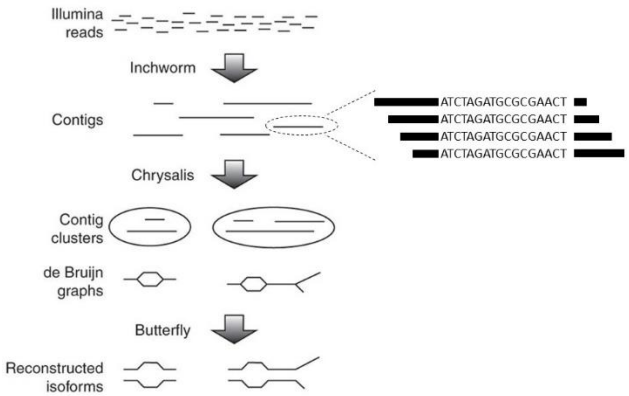


Figure 17 : Principe de l'assemblage *de novo* utilisé par l'outil Trinity (adaptée de [112]). Trinity utilise les données de séquençage pour d'abord construire les contigs. Puis l'outil regroupe les contigs en cluster puis en graphique de Bruijn. Puis l'outil détermine les séquences les plus probables.

Chez l'homme l'assemblage *de novo* présente peu d'intérêt étant donné que le génome et le transcriptome sont largement étudiés. Ainsi les fichiers de références du transcriptome ou du génome sont facilement accessibles. Par exemple, le projet GENCODE (https://www.gencodegenes.org/) propose en libre accès les fichiers Fasta des transcrits et du génome humain. L'alignement des *reads* de RNA-seq sur le transcriptome a l'avantage d'être rapide et peu avide en ressources informatiques. Par exemple, l'outil Bowtie [113] offre la possibilité d'aligner rapidement les *reads* sur un transcriptome de référence. Cependant ce type d'approche limite grandement la découverte de nouveaux transcrits car les *reads* issues du RNA-seq sont uniquement alignées sur les transcrits déjà décrits.

L'alignement sur le génome nécessite de « découper » le *read* en différents blocs, chacun d'entre eux représentant la séquence d'un exon sur le génome (Figure 18). Cette nécessité de découper le *read* en plusieurs fragments augmente le temps de calcul. Cependant elle permet la découverte de nouvelles jonctions d'épissage et donc de nouveaux transcrits. L'outil STAR [114] a notamment été développé dans cette optique. Brièvement, l'outil STAR procède en deux étapes, définir le bloc optimal ou *Maximal Mappable Prefix* (MMP) et regrouper ces MMPs pour calculer un score d'alignement. A partir de la première base du *read*, STAR définit les MMPs en prenant en compte la séquence du *read*, la position dans le génome et la séquence du génome. Les MMPs obtenus pour chaque *read* sont regroupés. Puis un score d'alignement est calculé en prenant en compte le nombre de MMPs, la distance entre eux et le nombre de disparités entre le MMP et le génome.

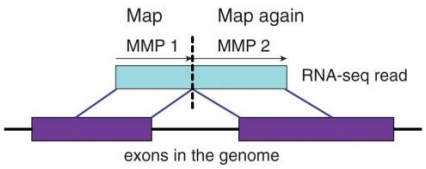


Figure 18 : Découpage d'un *read* issu du RNA-seq pour réaliser un alignement sur deux séquences exoniques (d'après [114]).

Indépendamment de la stratégie choisie pour l'alignement, un certain nombre de paramètres doit être défini par l'utilisateur pour assurer un alignement optimal. Parmi les principaux paramètres, sont retrouvés la qualité minimum de séquençage, le nombre de discordances et la taille minimale d'alignement ou de chevauchement des *reads*. La qualité de séquençage est dépendante de la plateforme de séquençage utilisée. Par exemple pour une plateforme type Illumina® le score PHRED minimal est de 30 tandis que la plateforme Oxford Nanopore® atteint un score PHRED de 10 à 20. Le nombre de disparité est le plus souvent défini par rapport à la longueur du *read*, par exemple pour un *read* de 100 nt, l'outil STAR préconise un nombre maximal de disparités à 8 nt. En complément de ces paramètres, il est également recommandé de réaliser un séquençage en *paired-end* avec une longueur de *read* d'au moins 100 nt pour optimiser l'alignement. En effet, un séquençage *paired-end* permet de corriger d'éventuelles erreurs de séquençage et la longueur des *reads* facilite la détection des événements d'épissage [115].

Suite à l'alignement, les séquences et coordonnées des *reads* sont enregistrées dans un fichier BAM, ce dernier étant la pierre angulaire pour la suite des analyses des données RNA-seq.

c. Identification des transcrits

Une fois les coordonnées génomiques déterminées durant l'alignement, les *reads* sont associés aux transcrits et le cas échéant à de nouveaux transcrits. Pour cela il est nécessaire de disposer d'un fichier d'annotation contenant les coordonnées des différents transcrits, le plus souvent au format BED ou GFF.

L'approche la plus simple consiste à associer les *reads* aux transcrits les plus proches. A ce titre, l'outil BEDtools [116] propose d'identifier les structures les plus proches des coordonnées des *reads*. Les structures peuvent être soient la position d'un gène, d'un transcrit ou d'un exon.

Cependant ce type d'approche ne permet pas d'associer les *reads* entre eux. En effet dans le cadre d'un séquençage *short read* chaque *read* ne supporte qu'une partie d'un transcrit. Aussi il a rapidement été développé des outils permettant de regrouper les différentes *reads* entre eux pour identifier l'ensemble du transcrit. Parmi ces outils, Cufflinks [117] est l'un des premiers à avoir été développé. Ce dernier regroupe, pour chaque gène, les *reads* dont les jonctions d'épissage sont compatibles entre elles. Puis il détermine les possibles isoformes en se basant sur le critère publié par Xing en 2004 [118]. Brièvement ce critère est basé sur la concordance entre les séquences des *reads* et les séquences des exons connus. Puis pour chaque isoforme possible, l'abondance des *reads* supportant ces isoformes est estimée. Cette estimation permet ensuite de vérifier la vraisemblance de chaque isoforme et de leur abondance au sein des tissus. Ainsi Cufflinks est capable de conjecturer les isoformes supportées par les données RNA-seq. Cependant ce type d'approche ne permet que de présupposer la véracité de ces isoformes. En effet, l'exacte reconstruction des transcrits à partir des données *short reads* est le plus souvent incomplète

[119]. Ceci explique en partie l'essor grandissant du séquençage *long read* en réponse à cette limitation technique.

d. Comptage des reads

Suite à l'alignement et à l'identification des *reads* issues du RNA-seq, un comptage du nombre de *read* par structure génétique est la dernière étape avant toute analyse d'expression. Habituellement le comptage se fait par gènes afin d'avoir, à l'instar des *microarrays*, un profil d'expression des gènes. L'outil HTSeq-count [120] a notamment été développé pour permettre ce comptage. Ce dernier nécessite le fichier d'alignement SAM et un fichier d'annotations au format GFF ou BED. Puis il génère en sortie un fichier texte indiquant le nombre de *reads* pour chaque gène présent dans le fichier d'annotation. Le comptage peut aussi se faire transcrit par transcrit, exon par exon et jonction par jonction. A titre d'exemple l'outil FeaturesCount [121] permet un comptage personnalisé des données. Cependant la nature du comptage (gène, transcrit, exon, jonction) doit être prise en compte pour l'analyse statistique.

2. Les outils biostatistiques

Les analyses statistiques des données RNA-seq ont pour but de mettre en avant une différence d'expression significative entre différents échantillons. Elles se décomposent habituellement en deux étapes : la normalisation du nombre de *reads* et l'ajustement des données à un modèle statistique pour calculer la différence d'expression entre les échantillons. Le plus souvent les outils proposés ont été mis au point pour déceler une différence significative d'expression de gènes à travers différentes conditions. En parallèle de ces analyses, plusieurs méthodes sont aussi proposées pour visualiser les données à chaque étape.

a. Visualisation des données brutes

Avant de procéder aux analyses statistiques, un certain nombre d'outils propose de visualiser les données brutes. Parmi ces outils, un des plus populaires est IGV (*Integrative Genome Viewer*) [122]. Il propose entre autre de visualiser les données juste après alignement (fichier BAM). La couverture des régions d'intérêt est représentée sous forme d'histogramme et les disparités entre le génome de référence et le *read* sont indiquées par un code couleur (Figure 19). Plus spécifiquement pour les données de RNA-seq, IGV propose aussi de représenter les jonctions d'épissage avec le nombre de *reads* supportant ces dernières par un graphique nommé *Sashimi plot* (Figure 20). Les *Sashimi plot*s ont d'abord été développés dans le cadre de l'outil MISO (*Mixture-of-Isoform*) [123] qui, à l'instar de l'outil Cufflinks, a pour but de déduire la structure des transcrits à partir de données RNA-seq. Ces *Sashimi plots*

représentent la couverture des exons sous forme d'histogramme et illustrent les *reads* à cheval entre deux exons par un arc de cercle avec le nombre de *read* supportant cette jonction d'épissage [124].

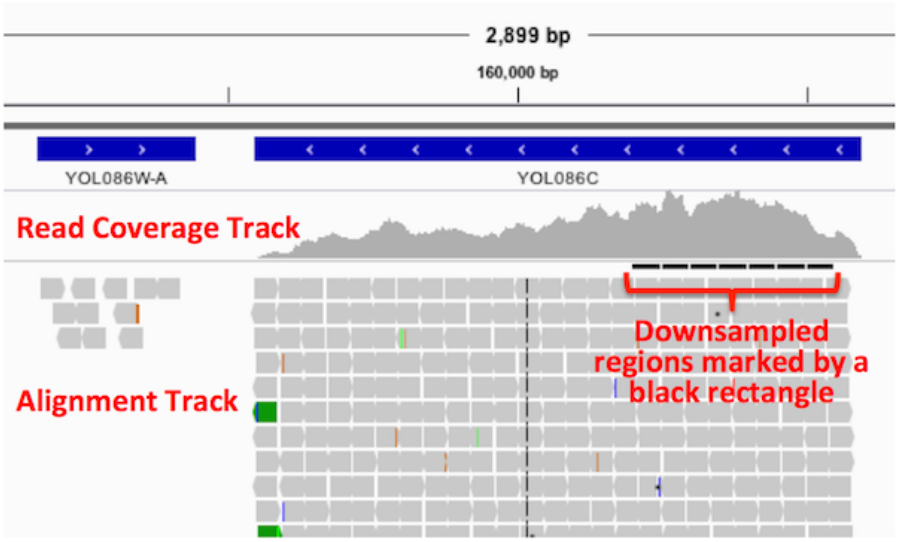


Figure 19 : Capture d'écran de la visualisation d'un fichier BAM par IGV. Les rectangles bleus correspondent aux structures de référence. L'histogramme au nombre de *reads* alignées sur chaque position. Chaque *read* est symbolisée par un rectangle gris en bas de l'écran.

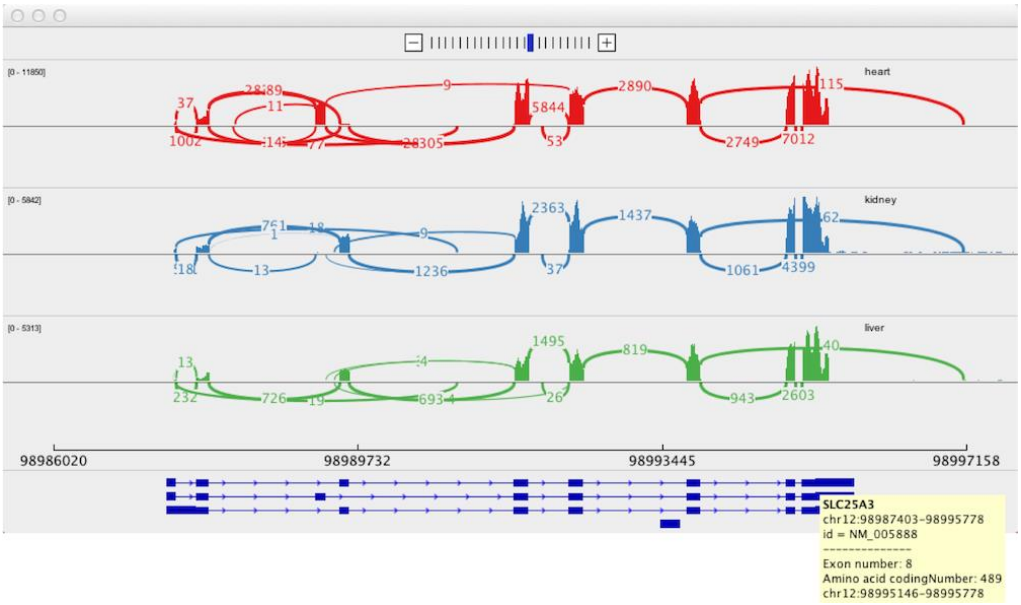


Figure 20 : Capture d'écran d'un *Sashimi plot* tracé par IGV. Les rectangles bleus correspondent à la structure des transcrits de référence. Sur les exons la couverture est représentée sous forme d'histogramme puis les jonctions par des arcs de cercle dont le chiffre correspond au nombre de *read* supportant cette jonction.

En complément de ces méthodes de visualisation, la couverture de séquençage peut aussi être représentée par le nombre de *read* par région d'intérêt *via* le *reads per KB per million reads* (RPKM).

Le RPKM représente le nombre de *read* d'une région d'intérêt (par exemple un exon) par rapport au nombre total de *read* et à la longueur du transcriptome, *via* la formule :

$$RPKM = \frac{10^9 \times C}{N \times L}$$

Où C est le nombre de *reads* de la région d'intérêt, N le nombre total de *read* et L la longueur du transcriptome en nt [125].

b. Normalisation du comptage de reads

Il est crucial de normaliser le comptage de *reads* avant de procéder à une analyse statistique. En effet un certain nombre de paramètres peut influencer artificiellement le nombre de *reads* par gène et donc conclure à tort une différence d'expression entre deux conditions [126]. Parmi ces paramètres, intervient la taille des librairies (le nombre de *reads*), la longueur du gène et le pourcentage en GC [127]. Si le RPKM prend en compte la taille des séquences couvertes, la plupart des méthodes de normalisation ne prennent en compte uniquement que le nombre global de *reads*. Ainsi la méthode la plus simple consiste à pondérer le comptage pour chaque région (ex : gène) et pour chaque échantillon, par le nombre total de *reads* de cet échantillon et la moyenne de couverture des échantillons, *via* l'équation :

$$\widehat{C_{ij}} = \frac{C_{ij}}{T_j} \times \bar{T}$$

Où $\widehat{C_{ij}}$ est le nombre de *reads* ajusté pour le i^{ème} gène du j^{ème} échantillon, C_{ij} est le nombre de *reads* observé pour le i^{ème} gène du j^{ème} échantillon, T_j le nombre total de *reads* du j^{ème} échantillon et $\bar{T}$ la moyenne du nombre total de *reads* parmi les j échantillons. Plusieurs variantes ont été proposées en remplaçant le compte total de *reads* par le 3ème quartile, ou la médiane [126].

Afin de réduire l'impact des valeurs extrêmes sur la normalisation du comptage. Il existe également une méthode de normalisation basée sur la moyenne géométrique du comptage pour chaque gène et échantillon. Ainsi un facteur de correction $\widehat{s_j}$ est estimé pour chaque j^{ème} échantillon. Ce dernier est la médiane des ratios entre la couverture de chaque gène au sein de l'échantillon et la moyenne géométrique de l'expression du gène parmi l'ensemble des échantillons, *via* l'équation :

$$\widehat{s_j} = mediane \frac{C_{ij}}{(\prod_{v=1}^{m} C_{iv})^{1/m}}$$

Où m est le nombre d'échantillons et $(\prod_{v=1}^{m} C_{iv})^{1/m}$ la moyenne géométrique du i^{ème} gène parmi les m échantillons [128]. Cette méthode d'ajustement est également proposée dans une version similaire par

Robinson et Oshlack [129]. Cette version, nommée *trimmed mean of M values* ou TMM, applique un filtre afin d'éliminer les valeurs extrêmes de couverture.

En ce qui concerne l'épissage, les méthodes de normalisation sont généralement basées sur le ratio entre le nombre de *reads* supportant un épissage donné et le nombre de *reads* supportant un second épissage au sein d'une même région. Ce calcul abouti à un pourcentage d'abondance relative entre deux épissages, le plus souvent noté psi ou Ψ [130]. Cette méthode a été adaptée aux données de RNA-seq d'abord pour les sauts d'exons au sein de l'outil MISO [123] puis généralisée à l'ensemble des épissages possibles [131].

c. Modélisation du comptage de reads

Avant de procéder à la modélisation des données, il est recommandé de réaliser une analyse non-supervisée. En effet ces analyses vont permettre d'isoler des groupes d'échantillons selon l'expression globale des gènes. Parmi ces approches non-supervisées, l'analyse en composante principale ou ACP est une méthode particulièrement utilisée. L'ACP consiste à réduire le nombre de valeurs sans perte notable d'informations afin de synthétiser cette masse d'informations sous une forme exploitable et compréhensible. Pour cela 2 ou 3 composantes principales sont calculées. Celles-ci sont des nouvelles variables combinant de manière linéaire l'expression des gènes. L'ajustement des composantes principales est obtenu par maximisation de la variance des projetées afin de réduire la perte d'information entre les différentes variables (Figure 21). Une autre approche non supervisée est la clustérisassions hiérarchique. Cette méthode consiste en un regroupement successif de points par ordre de proximité décroissante (Figure 22). Ainsi il est possible de regrouper les échantillons en fonction de leur profil d'expression.

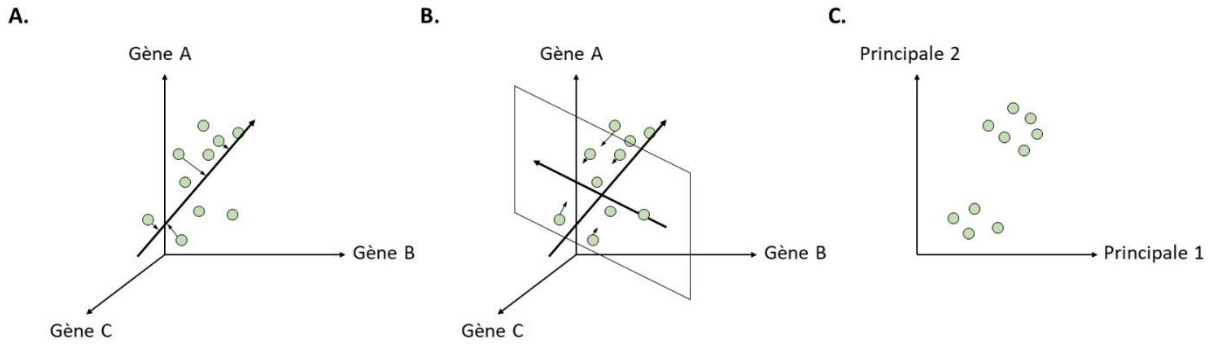


Figure 21 : Exemple ACP pour 10 échantillons avec l'expression de 3 gènes (A, B, C). **A.** Ajustement de la 1ere composante principale en maximisant la variance des projetées. **B.** Sur un plan orthogonal à la 1ere composante principale, ajustement de la seconde composante principale en maximisant la variance des projetées. **C.** Projection des données sur les deux composantes principales

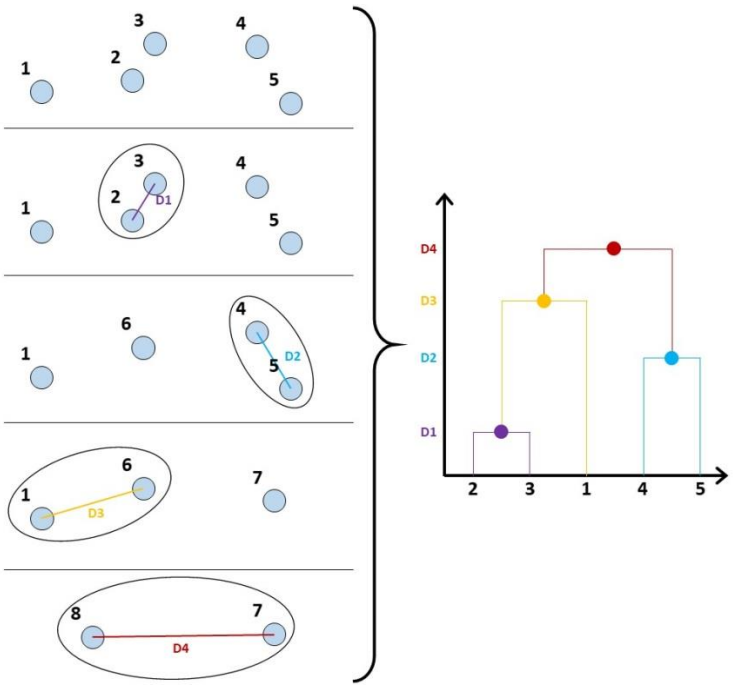


Figure 22 : Principe de la clustérisassions hiérarchique. Ici sont représentés 5 points (panel de gauche). D'abord la distance minimale entre ces points est mesurée puis reportée sur un graphique (panel de droite). Puis les deux points les plus proches sont regroupés en un barycentre. Ces étapes sont répétées jusqu'à ce qu'il ne reste plus que deux points.

Les analyses statistiques peuvent se diviser en deux groupes : les analyses paramétriques et non-paramétriques. Pour les analyses paramétriques, les données observées sont ajustées à un modèle mathématique préalablement défini, le plus souvent une loi de probabilité. Elle est un modèle mathématique qui pour une variable donnée permet de calculer la probabilité d'observer une valeur de cette variable. Dans le cadre du RNA-seq, la variable est l'expression des gènes dont les valeurs correspondent aux nombres de *reads* alignées. Ainsi entre deux conditions expérimentales, A et B, il est possible de calculer la probabilité (ou *p-value*) en fonction de l'expression des gènes. Cette probabilité permet ensuite de conclure sur une différence d'expression significative ou non. L'utilisation de loi de probabilité nécessite que les données soient ajustables à cette loi de probabilité et qu'elles soient suffisamment informatives pour permettre l'ajustement. Il est donc impératif de savoir quelle loi de probabilité peut le mieux décrire les données et de s'assurer d'avoir suffisamment d'observations pour permettre l'ajustement de celles-ci.

A l'opposé, les analyses non-paramétriques n'impliquent pas de loi de probabilité prédéfinie. Aussi les données peuvent être étudiées sans faire d'hypothèse sur une loi de probabilité et avec peu d'observations. En revanche, les analyses non-paramétriques ont en commun d'avoir une moindre puissance statistique que les analyses paramétriques. La puissance statistique est la probabilité de rejeter l'hypothèse nulle si l'hypothèse nulle est effectivement fausse. Parmi les analyses non-paramétriques

applicables au RNA-seq, le test exact de Fisher permet de mettre en évidence une différence d'expression avec peu de réplicas par condition. Cette méthode est intégrée à l'outil NOISeq [132].

Actuellement la plupart des outils utilisés pour étudier les données de RNA-seq sont basés sur des approches paramétriques. Initialement, les outils développés pour les *microarrays* ont été adaptés aux données RNA-seq tel que l'outil limma [133] basé sur l'utilisation d'un modèle linéaire continu. Cependant les données issues des *microarrays* sont continues car il s'agit d'intensité de fluorescence alors que les données de RNA-seq sont discrètes. En effet, il s'agit d'un comptage de *read* par gène. Les variables continues contiennent une infinité de valeurs au sein d'un intervalle, ex : la taille d'une personne. Les variables discrètes sont constituées de valeurs finies, ex : un lancer de dé. Ainsi, des outils statistiques reposant sur l'utilisation de lois de probabilités discrètes ont été par la suite développés spécifiquement pour le RNA-seq. Les lois de probabilités discrètes comprennent de manière non exhaustive les lois : uniforme, hypergéométrique, binomiale, de poisson et binomiale négative. Cependant les lois uniforme, hypergéométrique et binomiale ne peuvent s'appliquer aux données de RNA-seq. En effet, la loi uniforme présuppose une expression constante de tous les gènes. La loi hypergéométrique ne permet pas de modéliser un grand nombre de données. La loi binomiale tend à sous-évaluer la variabilité des données.

A l'inverse la loi de poisson, qui est une simplification de la loi binomiale, offre la possibilité de modéliser les évènements rares. En effet la probabilité qu'une séquence observée soit située sur la région d'intérêt parmi les millions de fragments lus par la plateforme de séquençage est très faible. La loi de poisson est une loi à un paramètre nommé λ correspondant à la fois à la moyenne et à la variance. Ainsi la probabilité d'observer k *reads* sur le i^{ème} gène du j^{ème} échantillon est calculée *via* l'équation :

$$P(X = k_{ij}) = \frac{\lambda_i^{k_{ij}}}{k_{ij}!} \times e^{-\lambda_i}$$

La loi de poisson est notamment utilisée par l'outil DEGSeq pour modéliser les données de RNA-seq [134]. Cependant, il persiste un problème de sur-dispersion. En effet, les données de comptage entre les gènes sont très hétérogènes et la variance souvent supérieure à la moyenne de l'expression de ces gènes. Aussi il a été proposé d'utiliser la loi binomiale négative. Cette loi modélise le nombre d'échecs pour obtenir *n* succès. Autrement dit, quel est le nombre de fragments à séquencer pour obtenir une couverture *n* du gène d'intérêt. Les paramètres de la loi sont la moyenne μ et la variance σ² définis *via* l'équation :

$$\begin{cases} \mu = \dfrac{n(1-p)}{p} \\ \sigma^2 = \dfrac{n(1-p)}{p^2} \end{cases} ; \forall p \in \{0;1\} \Rightarrow \sigma^2 > \mu$$

Où $(1-p)$ est la probabilité d'échec. Ainsi la variance d'une loi binomiale négative est toujours supérieure à sa moyenne, ce qui permet de prendre en compte la sur-dispersion des données de RNA-seq. Par conséquent cette loi de probabilité est utilisée dans de nombreux outils statistiques dédiés au RNA-seq : edgeR [135], DESeq [128], DESeq2 [136], baySeq [137], EBSeq [138], CuffDiff 2 [139]. Ces outils sont utilisés pour isoler des gènes différentiellement exprimés entre deux conditions expérimentales ou plus, en se basant sur la probabilité que ces gènes ne soient pas différentiellement exprimés. L'effectif par condition minimal recommandé pour un ajustement optimal est de 5 échantillons par condition [140]. Lorsque le nombre de *reads* sur un gène est particulièrement grand (> 10^3 – 10^4 *reads*), alors la variable du nombre de *reads* peut être assimilée à une variable continue. Dans ce cas de figure, la probabilité gamma est utilisée pour modéliser l'expression. C'est pourquoi certains outils intègrent, en plus de la loi discrète binomiale négative, la loi continue gamma.

En parallèle du calcul de probabilité, un autre paramètre est usuellement calculé pour représenter la différence d'expression, le *log2 fold-change*. Le *log2 fold-change* est le logarithme népérien en base 2 du rapport de l'expression entre deux conditions, soit pour un i^{ème} gène entre les conditions A et B :

$$\log_2 FC_i = \log_2\left(\frac{K_{A_i}}{K_{B_i}}\right) \Longrightarrow \begin{cases} \log_2 FC_i = 0 \Leftrightarrow K_{A_i} = K_{B_i} \\ \log_2 FC_i = 1 \Leftrightarrow K_{A_i} = 2K_{B_i} \\ \log_2 FC_i = 2 \Leftrightarrow K_{A_i} = 4K_{B_i} \end{cases}$$

Où K est le nombre de *reads* sur le gène. Un *log2 fold-change* est considéré comme élevé s'il est supérieur ou égale à 2. Le log2 fold-change est représenté graphiquement sous forme de *MA-plot* ou de *volcano plot* (Figure 23). Le *MA-plot* représente le *log2 fold-change* en fonction de l'expression des gènes. Le *volcano plot* trace la probabilité calculée par l'outil en fonction du *log2 fold-change*.

Il apparait également des outils statistiques développés pour identifier une différence d'expression non pas du gène mais des transcrits. Ainsi l'outil DEXSeq a été conçu pour mettre en avant une différence d'expression d'exons [141]. Il existe également des outils comme CuffDiff2 [142] qui ont pour but de quantifier les isoformes des transcrits. Cependant ces derniers outils sont limités de par la nature des données RNA-seq *short-read* (voir section *Identification des transcrits*).

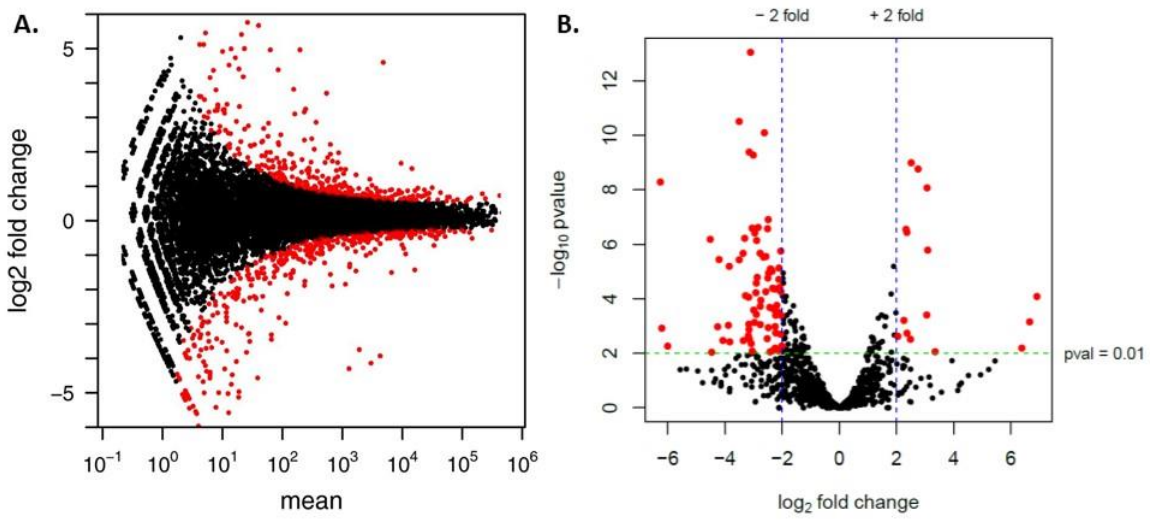


Figure 23 : Illustration de l'expression différentielle entre deux conditions. A. MA-plot, le log2 fold-change est représenté en fonction du nombre moyen de *reads*. B. Volcano plot, la probabilité calculée (p-value, ici en –log10) en fonction du log2 fold-change. Les points rouges représentent les gènes dont l'expression est significativement différente entre les deux conditions.

IV. Prédiction des défauts d'épissage

Avant l'ère du séquençage à haut débit, le nombre de variants détectés était relativement faible et ces variants pouvaient le cas échéant être testés au niveau de l'ARN. Cependant à partir des années 2000, l'amélioration et la réduction des coûts du séquençage à haut débit ont inversé cette tendance avec un accroissement constant de l'identification des variants au niveau ADN. Ainsi depuis les années 2000 de très grandes collections de variants sont établies. A titre d'exemple, le projet 1000 Genomes (https://www.internationalgenome.org/home) a réalisé le séquençage du génome complet pour 2 504 personnes, 84 millions de variants génétiques ont été identifiées [143]. De plus, 64 millions de ces variants sont rares (fréquence allélique < 0.5 %). En outre, les bases de données des variants étudiés à but diagnostic comprennent également un grand nombre de variants. ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) [144] et HGMD rapportaient respectivement plus de 450 000 variants et plus de 250 000 variants en août 2019. En raison de cette quantité de variants, les études fonctionnelles ne peuvent être réalisées sur tous ces variants. Aussi en marge de l'essor du séquençage à haut débit, un essor des outils de prédiction a été observé après les années 2000 (Figure 24). Ces derniers ont été développés afin de prioriser les variants pour des études fonctionnelles. En effet, ces outils se focalisent sur la prédiction d'une modification d'un ou plusieurs motifs d'épissage mais pas de l'impact de ces modifications sur les transcrits biologiques. Dans cette partie nous allons détailler de manière non exhaustive les principes de ces outils de prédiction.

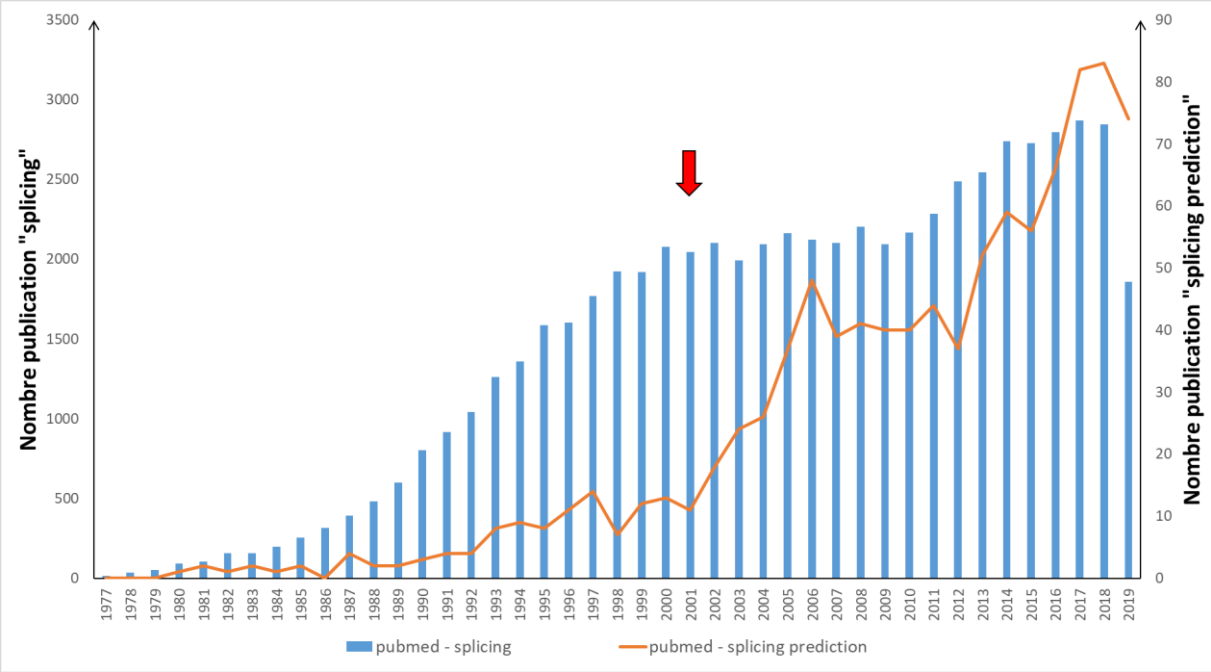


Figure 24 : Nombre d'articles référencés dans Pubmed comprenant les mots clés « *splicing* » et « *splicing prediction* » (juillet 2019). La flèche rouge marque le premier point d'inflexion en 2001 de la courbe du nombre d'article « *splicing prediction* »

1. Outils de prédiction dédiés aux sites d'épissage consensus

Les outils détaillés dans cette partie ont pour but de déceler les variants susceptibles de déstabiliser les motifs consensus d'épissage. Etant donné que les motifs consensus donneurs et accepteurs ont été les premiers à être identifiés et séquencés, ce sont ces motifs pour lesquels un score de prédiction a été attribué en premier. Parmi les différents outils proposés, les plus simples sont ceux reposant sur les matrices de position pondérée ou PWM pour *position weight matrix*. Une PWM est obtenue par alignement des séquences d'un motif donné. Puis à chaque position dans le motif, la fréquence des quatre nucléotides est calculée. Ainsi pour un motif long de 10 nt, la PWM sera un tableau de 4 lignes et de 10 colonnes. Le score *Splice Site Finder* ou SSF utilise par exemple la PWM [145]. Pour une séquence donnée, le score calcule la somme des fréquences à partir de la PWM. Puis il prend comme référence la séquence, la moins probable et la plus probable, c'est-à-dire les séquences dont la somme des fréquences est minimale ou maximale. Ces références sont ensuite utilisées pour calculer le score SSF de la séquence donnée (Figure 25). Un des principaux avantages de ce type de score est la simplicité de calcul sachant qu'il peut quasiment être calculé à la main. Ainsi ce type de calcul a été adapté aux autres motifs d'épissage (point de branchement et SREs) dans le cadre de l'outil *human splicing finder* ou HSF [146].

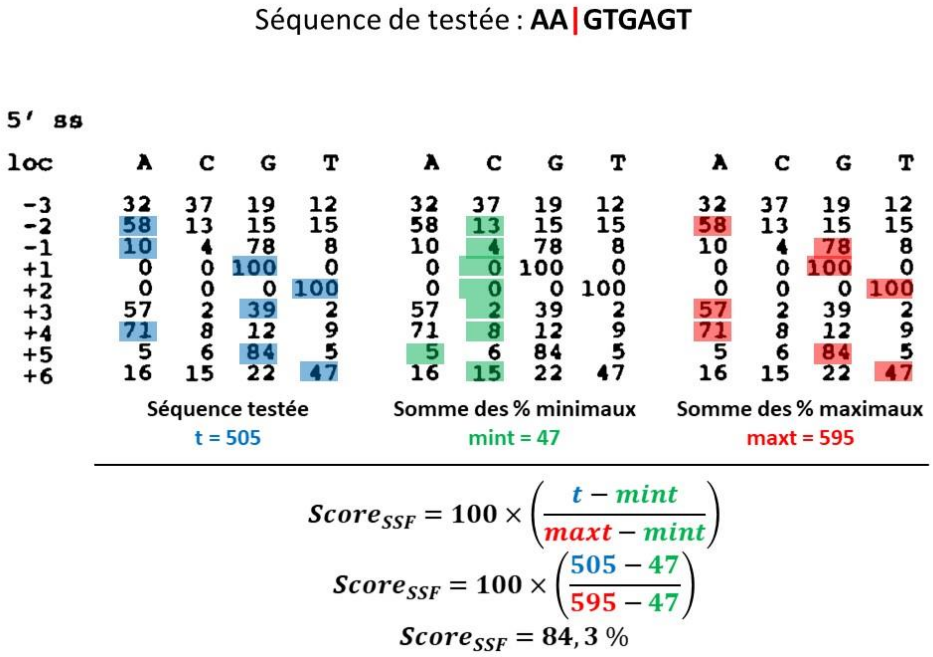


Figure 25 : Exemple de calcul du score SSF pour la séquence d'un site donneur AAGTGAGT. En bleu le score de la séquence testée, en vert le score minimal et en rouge le score maximal.

Cependant la principale limite des PWMs est le fait que les différents nucléotides composant le motif sont traités indépendamment. Aussi il a été proposé des outils dont les scores prennent en compte l'interaction des nucléotides de proche en proche. Pour cela des modèles probabilistes ont été utilisés

pour calculer les probabilités conditionnelles et ainsi prendre en compte l'interaction de proche en proche [147]. Exemple pour une séquence AG, la probabilité d'observer cette séquence $P(N_1N_2 = AG)$ peut se décomposer en :

$$P(N_1N_2 = AG) = P(N_1 = A) \times P_{N_1=A}(N_2 = G)$$

Où $P_{N_1=A}(N_2 = G)$ est la probabilité conditionnelle d'observer un G en deuxième position si le premier nucléotide est un A. Si les deux nucléotides sont indépendantes alors $P_{N_1=A}(N_2 = G) = P(N_2 = G)$ et si elles sont dépendantes alors $P_{N_1=A}(N_2 = G) \neq P(N_2 = G)$. Ainsi le calcul de la probabilité conditionnelle nous permet de savoir si l'apparition des nucléotides dans un motif est indépendante ou non. Un des outils à avoir particulièrement exploité ce concept est *MaxEntScan* (MES) [148]. Le modèle probabiliste utilisé par cet outil est la distribution du maximum d'entropie [149]. Ici l'entropie désigne la somme des probabilités des séquences possibles. Or dans le calcul de ces probabilités, un certain nombre de règles peut être fixé comme l'interaction des nucléotides de proche en proche ou bien l'interaction à un nucléotide d'écart ou plus. Puis sur un ensemble de séquence associé à un motif, le modèle tend à maximiser l'entropie, c'est-à-dire à accroitre les probabilités d'observer ces séquences, et inversement pour des séquences non associées au motif, de diminuer les probabilités. Ainsi en comparant la performance des différentes règles entre elles, les auteurs ont pu établir les interactions optimales pour les motifs des sites donneurs et accepteurs.

Il existe également des outils dont les scores ne sont issus ni d'une PWM, ni d'un modèle probabiliste mais d'étude ARN *in vitro*. C'est notamment le cas pour les motifs ESE/ESS, *via* les outils QUEPASA (*quantifying extensive phenotypic arrays from sequence arrays*) et ΔtESRseq (*delta of total ESRseq score change*) [24], [150]. Il a été testé en parallèle un grand nombre de séquences exoniques sur le pourcentage d'inclusion d'exons dans l'ARNm mature par MPRAs. Puis ce pourcentage d'inclusion est associé à chaque séquence de chaque motif. Cette valeur d'association est ensuite utilisée pour obtenir le score final d'une séquence exonique donnée.

2. Outils combinant plusieurs motifs d'épissage

Afin d'améliorer les prédictions d'épissage, de nombreux outils ne s'intéressent plus à un seul motif ou à un même groupe de motifs mais à une séquence ADN complète plus ou moins longue. Pour pouvoir apprécier l'adéquation d'une séquence test avec les motifs d'épissage, plusieurs méthodes ont été développées. En effet, chaque outil publié apporte une nouvelle pierre à l'édifice. Aussi dans cette partie nous ne pourrons pas décrire toutes ces méthodes. Cependant nous allons nous concentrer sur trois grandes types d'approches : les classifieurs non linéaires, le *machine learning* et le *deep learning*.

Parmi les classifieurs non linéaires, ceux couramment utilisés sont les *support vector machine* ou SVM, les arbres de décision et les *random forest* (RF). Le modèle SVM reprend un modèle développé durant les années 60 appelé perceptron [151]. Un perceptron est un modèle mathématique qui à partir d'un ensemble de *n* variables explicatives *X* renvoie une valeur *Y* bimodale : -1/+1. Cette dernière permet ainsi de séparer les données en deux groupes. Géométriquement, l'espace de dimension *n* est coupé en deux par une séparation appelé hyperplan (Figure 26A). Ainsi les modèles de type SVM calculent un hyperplan optimal afin de séparer au mieux le jeu de données. Pour définir cet hyperplan optimal, SVM replace les observations *X* dans un espace de dimension supérieur *Z*. Le plus souvent les variables *X* sont transformées par une fonction polynomiale. Cette espace Z est ensuite utilisé pour calculer l'hyperplan optimal (Figure 26B) [152]. Un exemple d'application de la méthode SVM est retrouvé dans l'outil SVM-BPfinder [153]. SVM-BPfinder permet la détection de points de branchement. Pour cela, l'outil combine avec un modèle SVM la séquence du motif point de branchement, la séquence du site consensus accepteur et la distance point de branchement-site accepteur.

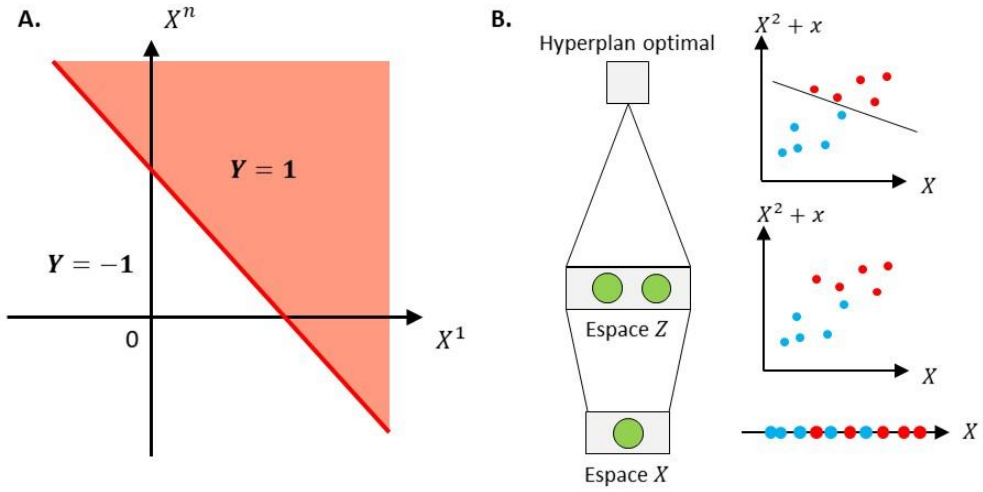


Figure 26 : Principe de l'analyse SVM. **A.** Illustration d'une séparation d'un espace à n dimensions séparé par un hyperplan issu d'un perceptron. **B.** Utilisation d'un espace Z de dimension supérieure à l'espace X pour obtenir l'hyperplan optimal.

Les arbres de décision sont des constructions automatiques visant à classer les données en deux groupes ou plus, selon un ensemble de variables explicatives. La racine de l'arbre correspond à la variable la plus explicative puis à chaque nœud de l'arbre un seuil décisionnel est appliqué pour basculer ou non sur la branche voisine [154]. Il existe différentes fonctions objectives permettant d'ajuster l'arbre décisionnel au mieux sur les données. A titre d'exemple l'outil *GeneSplicer* (GS) utilise un arbre décisionnel pour prédire l'utilisation de site donneur ou accepteur [155]. Cet arbre a été ajusté par la méthode nommée *maximal dependence decomposition* décrite par Burge et collaborateur [156].

Les *random forests* sont également basés sur le principe des arbres décisionnels. Cependant ici ce n'est pas un seul arbre qui est utilisé mais plusieurs, dont le nombre et la structure sont déterminés

automatiquement par l'algorithme [157]. L'intérêt de multiplier les arbres décisionnels est de stabiliser le modèle car la valeur renvoyée n'est plus unique mais multiple. Puis l'algorithme procède à un vote pour sélectionner la classe majoritaire [158] (Figure 27). Cette méthodologie a notamment été employée avec succès dans la reconnaissance des sites donneurs d'épissage [159].

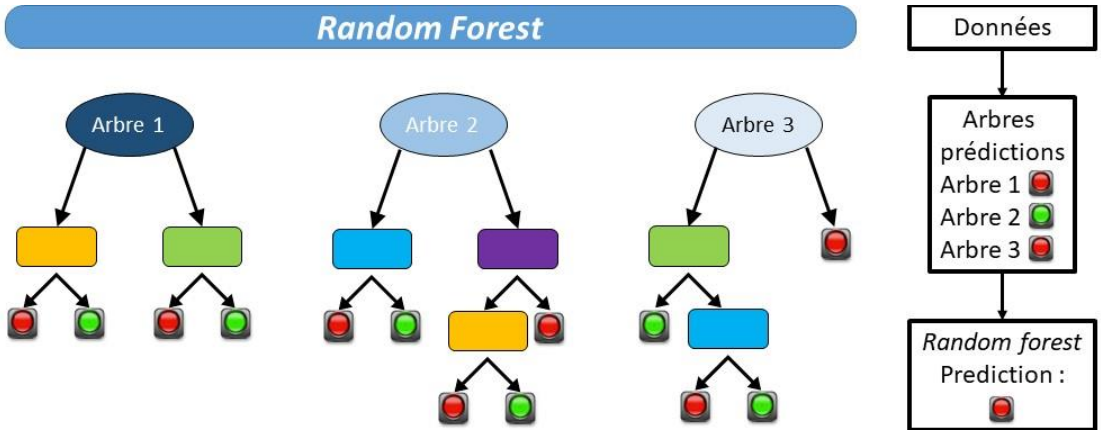


Figure 27 : Principe de l'algorithme random forest. L'algorithme est composé d'arbres décisionnels. Chacun d'entre eux prédit une classe. Puis la classe majoritaire correspond à la sortie de l'algorithme.

Le *machine learning* est un terme générique désignant un ensemble de méthodes statistiques. La plus représentée parmi elles est le réseau de neurones ou *neural network*. Cette méthode a été initialement développée à partir de 1985 par la communauté de l'intelligence artificielle [160], [161], son nom vient de l'homologie avec le cerveau humain. En effet, chaque « neurone » correspond à une fonction mathématique où le signal reçu par les dendrites est l'ensemble des variables explicatives *X* et le signal transmis par l'axone, la valeur renvoyée par la fonction. Puis ces fonctions mathématiques sont organisées de sorte que les sorties des premières fonctions servent d'entrée aux autres fonctions. Ainsi les *neural networks* sont habituellement représentés en couche (ou *layers*), où chaque couche contient plusieurs « neurones » (Figure 28). Chacune des fonctions utilisables est préalablement définie par l'utilisateur (régression linéaire, fonction trigonométrique, exponentielle, logarithme, …). Cette méthodologie a ainsi été appliquée par l'outil *Neural Network Splice* (NNS) [162]. NNS a été développé pour la reconnaissance des sites donneurs et accepteurs.

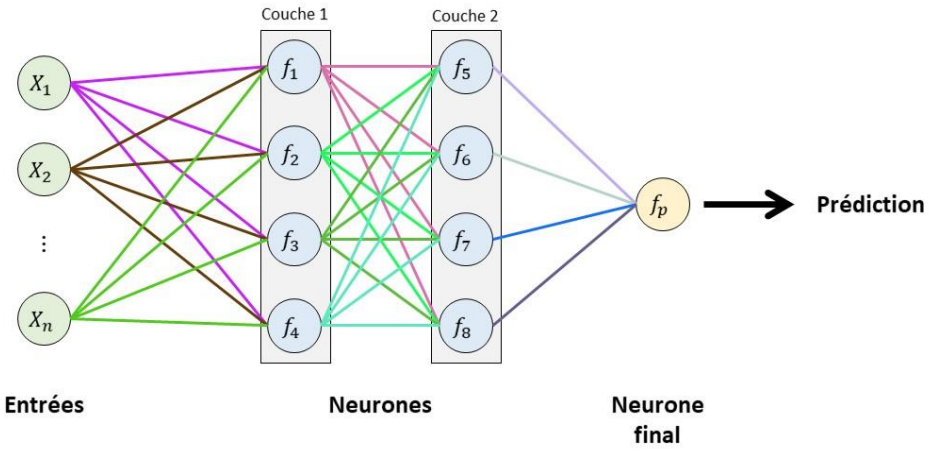


Figure 28 : Schéma général d'un neural network. Plusieurs neurones, chacun correspondant à une fonction mathématique, sont organisés en couche. Chacune des couches intègrent les données de la couche précédente puis ces sorties sont utilisées par la suivante. Pour aboutir un à neurone finale qui calcul le score final du modèle.

Le *deep learning* qui a connu son essor à la fin des années 2000, est une méthode mathématique reprenant le principe du *neural network* [163]. En effet, l'augmentation des capacités de calcul informatique et l'émergence de nouveaux algorithmes a permis de complexifier les réseaux neuronaux. Le *deep learning* se distingue des précédents *neural network* par deux points : la non nécessité de prédéfinir les fonctions utilisables (*i.e.* une plus grande autonomie du modèle) et utilisation d'un plus grand nombre de neurones et de *layers*. Aussi le *deep learning* peut s'appliquer à de plus larges domaines que l'initial *neural network*. Le principal objectif du *deep learning* est l'identification de motifs présents dans les données initiales, par exemple être capable de dire si une photographie montre un chien, une voiture, une maison, …. Pour cela l'outil procède à une transformation des données pour permettre une identification plus facile des éléments d'intérêt (Figure 29). Ainsi à chaque *layer* du *deep learning*, l'image est transformée afin de réduire les variables non pertinentes et d'exacerber les variations minimes des données d'intérêt [164], [165]. Appliqué à la biologie moléculaire, le *deep learning* est de plus en plus utilisé pour comprendre la structure des gènes. La prédiction des points de branchement a ainsi connu depuis peu un essor majeur d'outils fondés sur le *deep learning* : Branchpointer [166], LaBranchoR (*Long short-term memory network Branchpoint Retriever*) [167], RNABPS (*RNA branch point selection*) [168]. Mais, le *deep learning* est également utilisé pour la reconnaissance des sites d'épissage donneurs/accepteurs, comme par exemple avec l'outil SpliceRover [169].

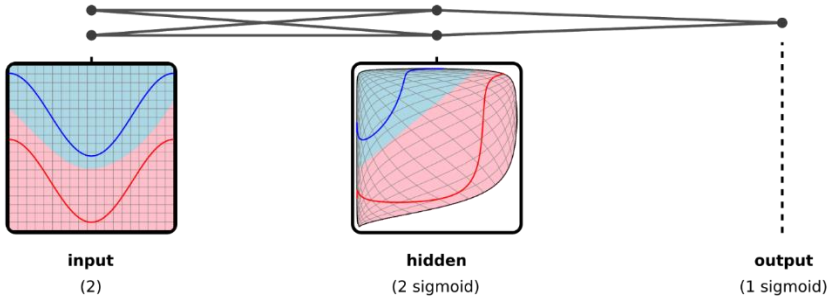


Figure 29 : Illustration de la transformation des données lors du deep learning pour avoir deux groupes linéairement séparables [170]. Ici deux groupes de données (bleu et rouge) sont illustrés. Puis le modèle retransforme les variables pour permettre une séparation linéaire des deux groupes.

3. Meta-scores

Les méta-scores sont des outils de prédiction intégrant plusieurs autres scores afin de valoriser l'information apportée par chacun. Cette approche permet notamment de calculer un score prenant en compte la variation d'autres scores entre la séquence sauvage et mutée. Parmi les méthodes employées pour combiner plusieurs scores, la plus simple est la régression logistique. Ce modèle permet de répartir les données en deux classes ou plus. Pour *p* variables *X* et *k* classes, le modèle combine linéairement les variables *X* par la fonction logit pour calculer la probabilité d'appartenir à la k^{ème} classe, *via* les équations :

$$logit\big(P_x(Y=k)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$P_x(Y=k) = \frac{e^{logit\left(P_x(Y=k)\right)}}{1 + e^{logit\left(P_x(Y=k)\right)}}$$

Où β_p sont les paramètres du modèle. Pour une classe à deux modalités (0/1, effet/pas d'effet, …), la première classe est prise comme référence et le modèle calcule la probabilité d'appartenir à la seconde classe. Par exemple, l'outil CRYP-SKIP intègre par régression logistique la variation des scores des motifs SREs pour prédire l'action d'un variant : soit un saut d'exon ou soit l'utilisation d'un site cryptique [171]. La classe de référence est le saut d'exon et le modèle calcule la probabilité de l'utilisation d'un site cryptique.

Les méthodes, précédemment décrites pour combiner plusieurs motifs d'épissage, sont également utilisées pour combiner plusieurs scores. Ainsi l'équipe de Xiaoming Liu a proposé en 2014 un nouvel outil de prédiction d'épissage pour les variants dans les régions consensus donneurs/accepteurs [172]. Cet outil utilise un algorithme *random forest* intégrant la variation des scores des motifs consensus d'épissage pour optimiser ces prédictions. L'outil SPANR (*Splicing-based Analysis of Variants*) qui prédit la variation du pourcentage d'inclusion d'un exon après action d'un variant, utilise le *deep*

learning [173]. En effet l'outil combine plusieurs centaines de variables dont la séquence des motifs et des scores de prédiction.

Ainsi nous pouvons constater une grande diversité d'outils de prédiction d'épissage aussi bien par leur nombre que par leurs approches, alors même que nous n'avons fait qu'effleurer ce sujet.

4. Evaluation des outils de prédiction

Face à cette diversité des scores de prédiction, l'utilisation de méthodes pour pouvoir les comparer entre eux est devenue critique pour la prédiction des défauts d'épissage. D'autant plus que l'accès à des données expérimentales peut s'avérer délicat. Par exemple, l'outil SVM-BPfinder a été entrainé sur plus de 100 000 introns pour localiser le point de branchement. Mais seulement 35 introns avec au moins un point de branchement expérimentalement prouvé ont été utilisés pour évaluer l'outil [153]. De même s'il existe plusieurs bases de données de variants (ClinVar, HGMD, …), il n'existe pas à ce jour de base de données régulièrement actualisée et dédiée aux études ARN *in vitro* de variants. Nous pouvons citer la base de données DBASS (*database of alternative splice site*) (http://www.dbass.org.uk/) [174], [175]. Mais DBASS ne contient que les données de la littérature sur les variants entrainant l'utilisation d'un nouveau site d'épissage et n'est plus maintenue depuis le début des années 2010. Ainsi les auteurs de ces outils de prédiction ont été contraints de générer leur propre jeu de données expérimentales pour les évaluer. Ces données sont issues soit de leur propre expérimentation, le plus souvent d'une approche artificielle de type MPRAs, soit de colliger les données de la littérature. Avec deux principales limites, les données artificielles ne correspondent pas nécessairement aux conditions réelles d'études des défauts d'épissage (technique *in vitro*, tissu-spécificité, …). Les données de la littérature sont sujettes à un biais de publication où le plus souvent seuls les résultats positifs sont rapportés. Ainsi, un jeu de données d'évaluation cohérent et représentatif des conditions réelles peut déjà s'avérer être un premier défi.

Pour l'évaluation des outils, deux cas de figure sont à considérer selon que les données soient discrètes ou continues. Une observation discrète est par exemple la reconnaissance ou non du motif d'épissage par le splicéosome ou bien l'altération ou non de l'épissage par un variant. Les données continues correspondent le plus souvent à l'expression relative d'un épissage alternatif (ex : $\Delta\Psi$) [173]. Cependant, les données continues sont moins fréquemment utilisées car étroitement dépendantes de la méthode employée pour les obtenir. Pour les données discrètes le calcul des critères d'évaluation des outils passe par l'édition d'un tableau de contingence. Ce tableau représente les effectifs observés en fonction de la nature des prédictions positives ou négatives (Figure 30). Le premier critère calculable à partir de ce tableau de contingence est l'exactitude. L'exactitude est la proportion de données correctement prédites

parmi l'ensemble des observations. Puis les autres principaux critères sont des probabilités conditionnelles estimées à partir du tableau de contingence avec :

- La sensibilité, aussi appelée *recall*, puissance ou *true positive rate* (TPR), est la probabilité d'avoir une prédiction positive si l'observation est positive.
- La spécificité, aussi appelée sélectivité ou *true negative rate* (TNR), est la probabilité d'avoir une prédiction négative si l'observation est négative.
- La valeur prédictive positive (VPP), aussi appelée précision, est la probabilité d'avoir une observation positive si la prédiction est positive.
- La valeur prédictive négative (VPN) est la probabilité d'avoir une observation négative si la prédiction est négative.

La sensibilité et la spécificité ont la particularité, contrairement aux VPP et VPN de prendre le problème à l'envers si l'on se place dans la situation d'un biologiste ayant identifié par NGS un variant et cherchant à évaluer par un outil son impact au niveau de l'épissage. En effet, la sensibilité et la spécificité calculent la probabilité d'avoir une prédiction correcte sachant la nature de l'observation. Cependant, elles sont peu sensibles à la répartition des données positives/négatives. Les VPP et VPN sont bien représentatives des conditions d'utilisation de l'outil. Cependant, elles ne sont relevantes que si la répartition positive/négative des données correspond à la répartition naturelle des données.

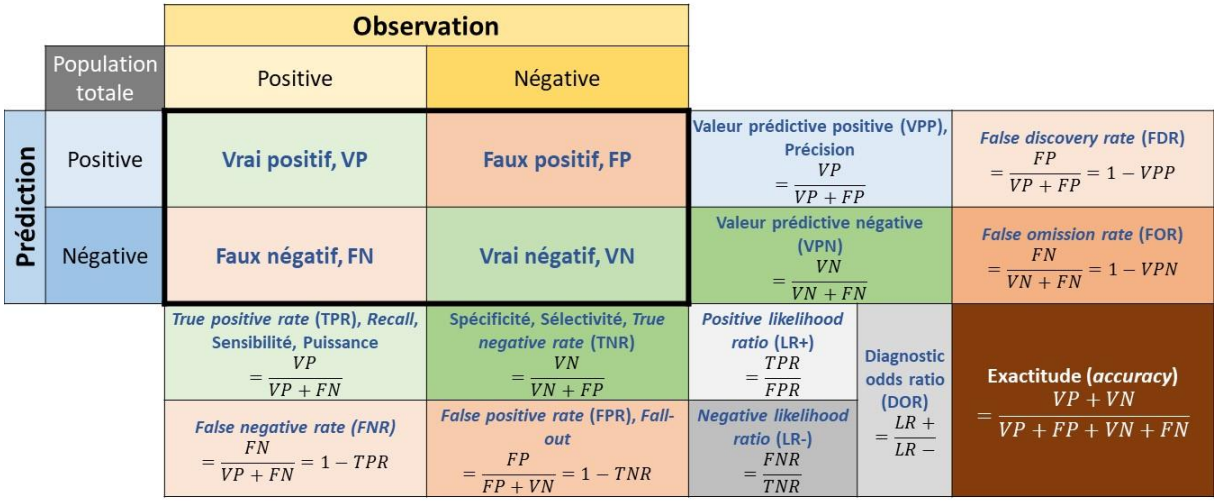


Figure 30 : Tableau de contingence (cf. encart) avec les critères utilisés pour l'évaluation des outils de prédiction.

La plupart des outils renvoyant un score continue et la définition d'une prédiction négative ou positive nécessite d'établir un seuil décisionnel. La courbe ROC (*receiver operating characteristic*) permet d'éviter l'utilisation d'un seuil arbitraire ou non-optimal [176]. En effet pour chaque valeur de seuil possible, la sensibilité et le *false positive rate* (FPR) (1-spécificité) sont calculés puis tracés pour obtenir cette courbe ROC (Figure 31). En plus de pouvoir définir un seuil optimal la courbe ROC permet de

calculer un autre critère d'évaluation des outils. L'aire sous la courbe ROC, appelée AUC (*area under the curve*), permet de représenter la performance de l'outil indépendamment de la valeur du seuil. La valeur maximale de l'AUC est 1, ainsi l'outil prédit parfaitement l'évènement. La valeur minimale est 0.5, autrement dit l'outil n'est pas plus performant qu'une sélection aléatoire des données. Le seuil optimal correspond aux valeurs maximales de sensibilité et de spécificité, c'est-à-dire au point d'inflexion de la courbe ROC. La sélection du seuil optimal se fait ainsi en minimisant l'écart entre la sensibilité et la spécificité. Pour sélectionner le seuil optimal, il est également possible d'utiliser le score de Youden, calculé *via* l'équation :

$$Youden = Sensibilité + Spécificité - 1$$

Pour les données continues, la corrélation est testée entre l'expression (données observées) et le score (données prédites). Cette mesure de corrélation est basée sur le calcul de la covariance et est utilisée pour calculer le coefficient de corrélation linéaire, ou coefficient de Pearson (r). La valeur de ce coefficient est représentée au carré (r^2) afin de s'affranchir du signe de la corrélation et varie de 0 à 1. Ainsi une valeur de 1 signifie que 100% de la variance de l'expression est expliquée par le score de l'outil. A l'inverse une valeur de 0 révèle l'absence de corrélation entre les deux variables. En cas de corrélation non linéaire entre les deux variables, il est utilisé le coefficient de Spearman, qui ne mesure pas la covariance sur les valeurs mais sur leur rang.

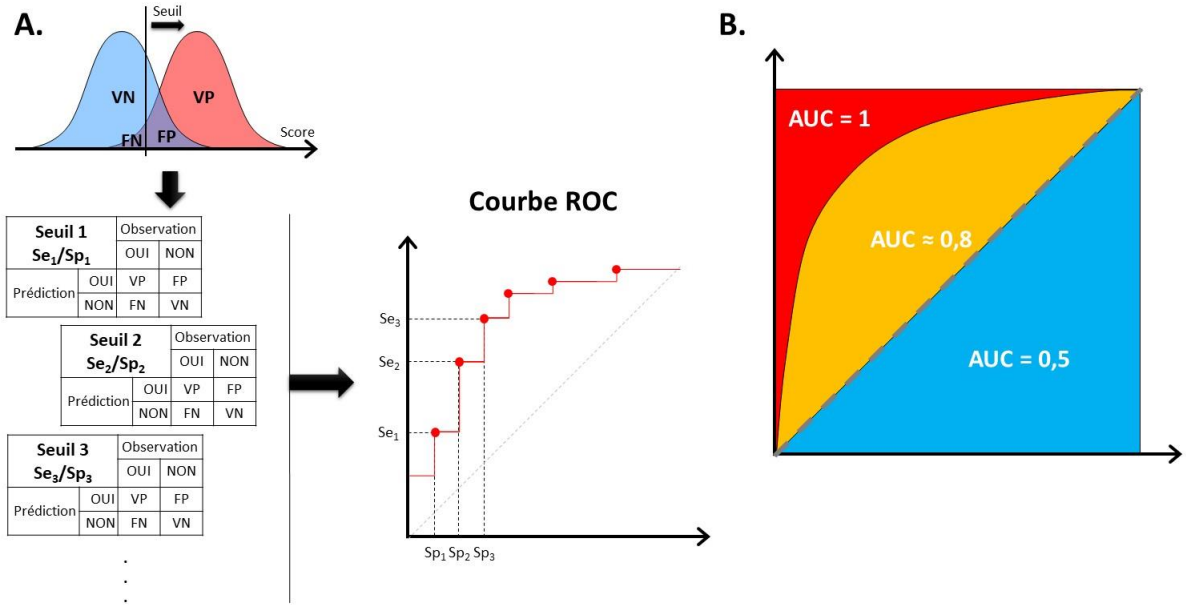


Figure 31 : Principe des courbes ROC. **A.** Pour chaque seuil possible, la sensibilité (Se) et la spécificité (Sp) sont calculées puis représenté graphiquement. **B.** Différentes valeurs d'AUC et leurs représentations graphiques.

En marge de ces critères d'évaluation mathématique, les outils peuvent également être évalués sur leur accessibilité et leur utilisation. En effet, un certain nombre d'outils, notamment ceux publiés depuis

plusieurs années, ne sont plus maintenus voir inaccessibles. Par exemple l'outil ASSP (*alternative splice site predictor*), publié en 2005, est actuellement inaccessible [177]. En ce qui concerne l'utilisation des outils, en absence de recommandations, nous pouvons observer une très grande diversité d'accès de ces outils. Certains d'entre eux nécessitent de générer les séquences ADN des régions d'intérêt, le plus souvent en format Fasta (ex : MES, GS ou SVM-BPfinder). L'outil devient alors moins facile à utiliser et augmente le risque de mésusage, notamment en cas d'étude de variants où l'utilisateur doit définir la séquence sauvage et mutée. C'est pour cela que de plus en plus d'outils proposent une solution *genome browser*. C'est-à-dire que l'outil requière la position de la région d'intérêt puis extrait lui-même la séquence ADN et si nécessaire la séquence mutée (ex : HSF, Branchpointer ou SpliceRover). Ces derniers peuvent soit utiliser des fichiers selon leur propre format ou bien sous un format plus standardisé comme le format VCF. La plupart des outils de prédiction sont accessibles en ligne et/ou en téléchargement, cette dernière option étant préférable pour traiter un grand nombre de données [178]. Que ce soit par la complexité de l'outil ou bien par son temps de calcul, certains auteurs proposent aussi une base de données avec les scores pré-calculés et utilisables par certains logiciels. Par exemple les bases de données dbscSNV [172], [179] ou SPIDEX [173] sont accessibles par le logiciel ANNOVAR (*Annotate Variation*) [180]. LaBranchoR propose une liste de points de branchement prédits accessibles par l'UCSC *Genome Browser* [108].

Ainsi l'ensemble de ces méthodes permet de définir un outil optimal pour prédire un défaut d'épissage.

V. Prédisposition aux cancers du sein et de l'ovaire : un modèle d'étude des variants splicéogéniques

Chez la femme, le cancer du sein est à la fois le plus fréquent avec la plus forte mortalité. En France, le cancer du sein tue plus de 12 000 femmes chaque année. Même si cette mortalité tend à diminuer, la prévalence, ne cesse de croître depuis les années 90. En 1990, 30 000 nouveaux cas étaient rapportés, puis en 2018, nous observons 58 400 cas annuels, soit +1,1 % par an en moyenne [181]. Le cancer de l'ovaire, bien que moins fréquent que le cancer du sein avec 4 700 nouveau cas en 2017, est responsable de la mort de 3 100 femmes chaque année [182]. Ainsi le cancer de l'ovaire est le cancer gynécologique avec le plus mauvais pronostic [183].

Environ 5 à 10 % des cancers du sein et de l'ovaire surviennent dans un contexte héréditaire [184]. Ces cancers héréditaires, aussi appelés syndrome HBOC (*hereditary breast and ovarian cancer*), suivent un schéma de transmission autosomique dominante. Ce fut en 1994 et en 1995 que les premiers gènes porteurs de cette hérédité ont été identifiés : *BReast CAncer 1* (*BRCA1*) et *BReast CAncer 2* (*BRCA2*) [185], [186]. En effet, la présence de variants entrainant la perte de fonction des gènes *BRCA1* et/ou *BRCA2* est associée à un risque accru de syndrome HBOC. Le risque cumulé à l'âge de 80 ans pour une femme de développer un cancer du sein est de 72 % pour le gène *BRCA1* et de 69 % pour le gène *BRCA2*. Quant au risque cumulé à l'âge de 80 ans pour le cancer de l'ovaire, il est de 44 % pour le gène *BRCA1* et de 17 % pour le gène *BRCA2* [187].

L'identification de ces variants relève d'un intérêt clinique majeur. La découverte chez un patient d'un variant délétère *BRCA1/BRCA2* permet d'adapter sa prise en charge clinique et thérapeutique. Elle permet aussi de rechercher ce variant chez les autres membres de la famille, afin de proposer à ceux porteurs de ce variant une surveillance appropriée et si nécessaire une chirurgie prophylactique des seins et des ovaires [188]. Elle permet également chez les non-porteurs de ce variant de lever toute surveillance spécifique au syndrome HBOC. En effet le risque de développer un cancer du sein ou de l'ovaire rejoint celui de la population générale. Ainsi, seulement en France, plus de 21 000 personnes ont été reçues en consultation d'oncogénétique pour la recherche de ces variants en 2017 [189].

Il en résulte qu'après plus de 20 ans d'études des variants *BRCA1/BRCA2* et de l'incidence du syndrome HBOC, ces derniers constituent une part importante des variants rapportés en génétique humaine. La base de données ClinVar rapporte actuellement des variants humains dans plus de 6 000 gènes mais les gènes associés au syndrome HBOC constituent la plus large fraction de ces variants (Figure 32). Par conséquent le syndrome HBOC constitue un bon modèle d'étude de variants pathogènes notamment au niveau de l'épissage.

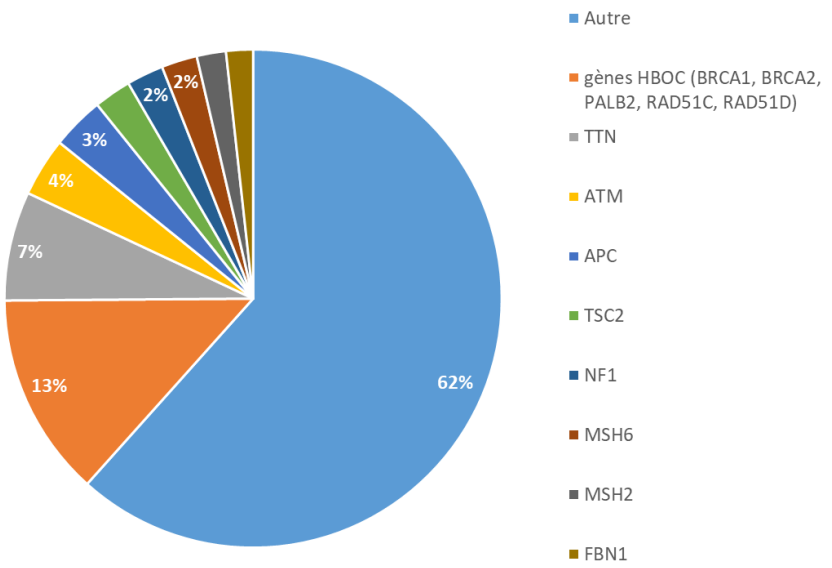


Figure 32 : Répartition du nombre de variants parmi les 100 gènes les plus représentés dans la base de données ClinVar (août 2019).

1. Gènes impliqués dans le syndrome HBOC

a. Gènes BRCA1 et BRCA2

Le gène *BRCA1* est situé sur le chromosome 17. Il est composé de 23 exons, dont 22 codants, distribués sur plus de 80 kb. La transcription de ce gène produit un transcrit principal de 7.2 kb qui est ensuite traduit en une protéine de 1 863 acides aminés. Le gène *BRCA2* est situé sur le chromosome 13. Il est composé de 27 exons, dont 26 codants, distribués sur environ 84 kb. La transcription de ce gène produit un transcrit principal de 11.4 kb qui est ensuite traduit en une protéine de grande taille avec 3 418 acides aminés. A l'instar de la majorité des gènes humains, *BRCA1* et *BRCA2* présentent plusieurs épissages alternatifs [131], [190], [191]. Les rôles biologiques de ces derniers restent encore à préciser. En effet les protéines BRCA1 et BRCA2 ont de nombreuses fonctions au sein de la cellule, notamment la recombinaison homologue de l'ADN. Mais nous pouvons aussi citer la régulation de l'apoptose, du cycle cellulaire, ou bien de l'ubiquitination.

La recombinaison homologue de l'ADN est un des principaux mécanismes utilisés par la cellule pour réparer les cassures double brins de l'ADN [192]. Brièvement, la coupure double brins active la protéine ATM (*ataxia telangiectasia mutated*) [193]. En parallèle se forme le complexe 53BP1. Ce dernier entraîne le recrutement d'une nucléase pour que l'un des brins de la partie en 3' soit dégradé ainsi que la protéine BRCA1 associée à BARD1 (*BRCA1 associated ring domain 1*). Ces deux dernières vont alors former un complexe avec PALB2 (*partener and localizer of BRCA2*) et BRCA2. Le brin d'ADN 3' non dégradé se voit alors entouré par la protéine RAD51 (*RAD51 recombinase*). Puis, ce brin s'hybride avec le brin complémentaire du second brin d'ADN. Ainsi le brin manquant est néo-synthétisé puis ligué au brin d'ADN initialement tronqué (Figure 33).

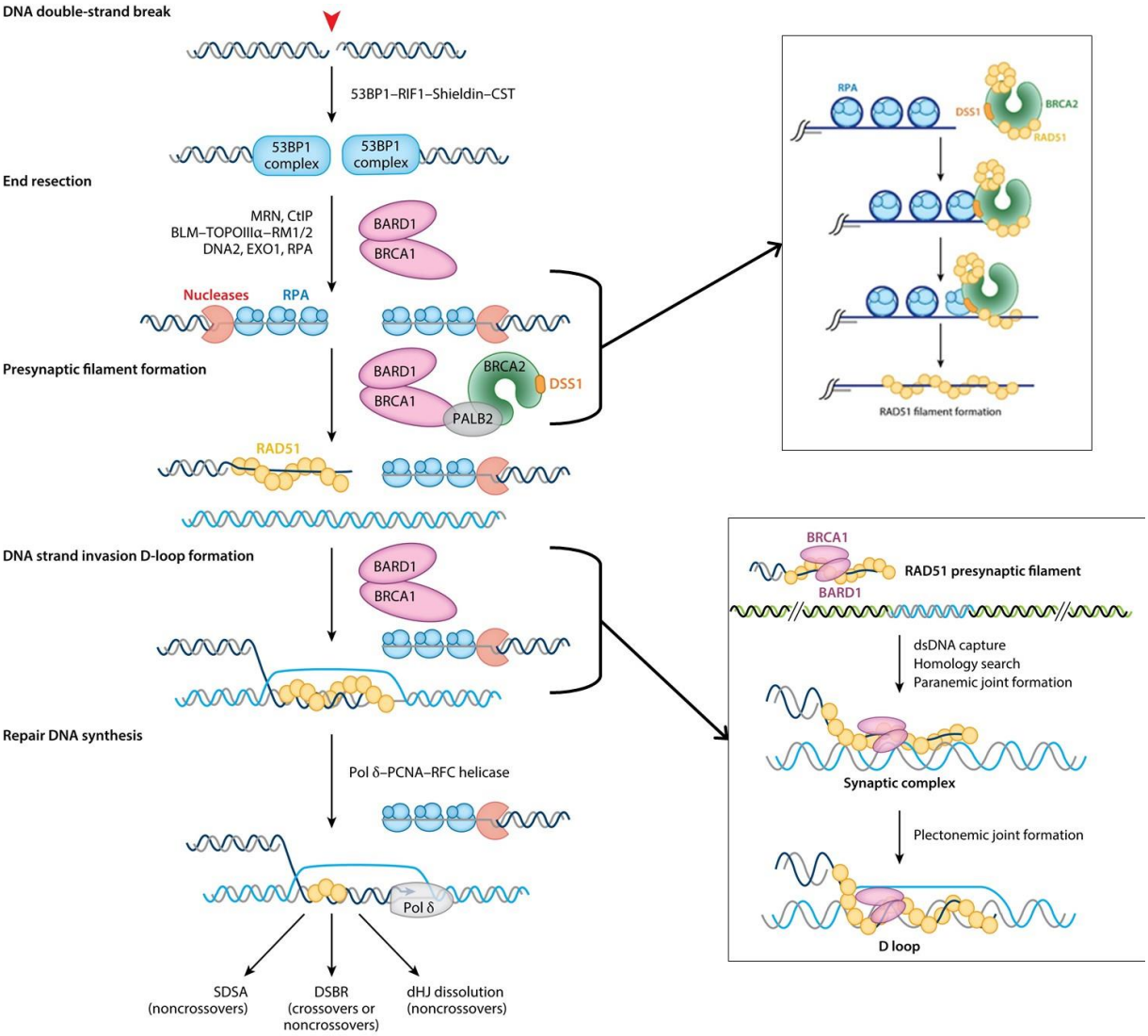


Figure 33 : Processus de la recombinaison homologue avec les principaux partenaires impliqués (adaptée de [194]). L'accent est surtout mis sur les protéines BRCA1/BRCA2 et leurs partenaires (DSS1–BRCA2–PALB2–BRCA1–BARD1). Abbreviations: dHJ, *double Holliday junction*; DSBR, *double-strand break repair*; SDSA, *synthesis-dependent DNA strand annealing*.

BRCA1 comprend un domaine catalytique RING (*Really interesting new gene*) en N-terminal. Ce domaine interagit avec la protéine BARD1 pour former un complexe hétérodimérique à activité ubiquitine ligase E3. Cette interaction BRCA1/BARD1 est d'ailleurs essentielle pour la stabilité des deux protéines [195]. Les enzymes à activité ubiquitine ligase E3 sont responsables de l'ubiquitination d'un grand nombre de protéines, y compris les histones et la protéine BRCA1 elle-même [195]. Cette ubiquitination permet de marquer les protéines à dégrader par le complexe protéasome. En plus de son activité ubiquitine ligase E3, l'interaction BRCA1/BARD1 est importante pour le rôle de BRCA1 dans la suppression de tumeurs et la réparation de l'ADN [196]. BRCA1 présente aussi deux domaines répétés en tandem BRCT (*BRCA1 Carboxy-Terminal*) sur la partie C-terminal. Ces domaines ont la capacité de

s'associer avec les facteurs de réponse aux dommages de l'ADN [197]. Immédiatement avant les domaines BRCT, il y a un domaine de fixation à la protéine PALB2 [198]–[200]. La région centrale de BRCA1, qui est codé par l'exon 11, comptant pour plus de 60 % de la protéine, est supposée participer à la structure secondaire de la protéine [197]. Elle participe aussi à la fixation sur l'ADN et de RAD51 [201].

La partie N-terminale de BRCA2 est impliquée dans la formation du complexe avec PALB2 [200]. Il y a huit domaines répétés BRC (*breast cancer*) (chacun faisant ≈ 30 résidus) qui interagissent avec RAD51 individuellement [202], [203], plus un domaine indépendant en C-terminal se liant aussi à RAD51, nommé CTRB (*C-terminal RAD51-binding*) [204]. BRCA2 possède aussi deux domaines de fixation à l'ADN, intitulés DBD (*DNA binding domain*) [205] et un autre N-terminal nommé NTD (*N-terminal DNA binding domain*) [206], [207].

b. Les gènes non-BRCA impliqués dans le syndrome HBOC

Bien que l'altération des gènes *BRCA1/BRCA2* soient étroitement associée au syndrome HBOC, les variants délétères de *BRCA1/BRCA2* ne sont identifiés que dans 10 à 20 % des suspicions de syndrome HBOC [189], [208]. L'hérédité manquante fait suspecter la présence de variants délétères dans d'autres gènes impliqués dans ce syndrome. En effet d'autres gènes précédemment décrits en oncogénétique sont associés à un risque augmenté de cancer du sein : *TP53* (*tumor protein p53*), *PTEN* (*phosphatase and tensin homolog*), *STK11* (*serine threonine kinase 11*) et *CDH1* (*cadherin 1*), responsables respectivement du syndrome de Li-Fraumeni, du syndrome de Cowden, du syndrome de Peutz-Jeghers et du syndrome de cancer gastrique héréditaire diffus [209]–[212]. Au cours de ces dernières années, il a également été identifié des gènes directement associés au syndrome HBOC. Ainsi, les variants délétères du gène *PALB2* se sont avérés être associés à un sur risque de cancer du sein, estimé à 35 % à 70 ans [213], [214]. Le gène *RAD51* s'est également montré être la cible de variants pathogènes dans le syndrome HBOC, en particulier les paralogues C (*RAD51C*) et D (*RAD51D*) [215], [216]. Les variants RAD51C/D sont majoritairement associés aux cancers de l'ovaire.

PALB2 est composé de 13 exons codants pour un transcrit de 3,6 kb qui est ensuite traduit en une protéine de 1 186 acides aminés. Le gène *PALB2* est situé sur le chromosome 16. La protéine PALB2 possède un domaine N-terminale se liant à BRCA1 [217]. Dans sa partie centrale, PALB2 présente aussi un site de liaison à la chromatine, ChAM (*chromatin-association motif*), facilitant les fonctions de PALB2 dans la réparation de l'ADN [218]. Puis, dans sa partie C-terminal, PALB2 est dotée d'un domaine WD40, riche en dipeptide tryptophane et acide aspartique [219]. Le domaine WD40 permet la fixation de BRCA2 sur PALB2 [220].

Il existe six paralogues de RAD51 nommés *RAD51B, RAD51C, RAD51D, XRCC2, XRCC3* et *SWS1AP1* [221]. Ces paralogues, bien que n'assurant pas la fonction recombinase de RAD51, régulent positivement la formation des filaments de RAD51 autour de l'ADN [222].

2. Interprétation des variants

L'interprétation des variants, dans le cadre d'un diagnostic moléculaire, peut se révéler particulièrement complexe. Dans le cadre du syndrome HBOC, les variants pathogènes sont ceux entrainant la perte de fonction/expression des gènes *BRCA1, BRCA2, PALB2* et *RAD51C/D*. Un variant peut-être délétère par les mécanismes suivants (liste non exhaustive) :

- Apparition d'un PTC : changement d'un codon en un codon stop (variant non-sens) ou décalage du cadre de lecture (variant hors phase) par des petites insertions/délétions.
- Altération d'un domaine fonctionnel : variant entraînant un changement d'acide aminé (variant faux-sens).
- Perte du codon d'initiation ou du codon stop.
- Réarrangement de grande taille (RGT) lié à des variants structuraux de l'ADN, comprenant entre autre :
    - Inversion partielle ou complète du gène.
    - Translocations, ex : insertion aléatoire d'une grande séquence ADN d'un autre chromosome.
    - CNV (*copy number variation*) correspondant à la délétion ou à la duplication du gène ou d'une partie du gène.

Pour certains variants, leur seule nature ne permet pas de conclure formellement sur leur impact protéique sans preuve complémentaire. Ainsi 5 classes de pathogénicité ont été proposées pour les pathologies cancéreuses héréditaires fréquentes par le groupe GCS (*Genetic Cancer Susceptibility*), ce dernier faisant partie de l'IARC (*International Agency for Research on Cancer*) [223]. Les variants de classes 1 et 2, variants bénins ou probablement bénins, comprennent les variants non pathogéniques. Les variants de classes 4 et 5, variants probablement pathogènes ou pathogènes, regroupent les variants délétères. Puis, la classe 3 correspond aux variants dont la pathogénicité ne peut être clairement démontrée. Ces derniers sont nommés VUS ou *variant of unknown significance* et ne sont pas utilisables pour le conseil génétique. Ces classes sont définies par la probabilité que le variant soit pathogène. Cette probabilité est calculée *a posteriori* sachant les données associées à ce variant, aussi notée $P_D(V)$ où V correspond à l'évènement variant pathogène et D aux données observées. Ces données sont habituellement la co-ségrégation du variant avec la pathologie au sein de la famille, la cooccurrence en trans avec d'autres variants pour un même patient, *etc*.

Grâce au théorème de Bayes, cette probabilité conditionnelle peut se décomposer en :

$$P_D(V) = \frac{P_V(D) \times P(V)}{P(D)}$$

Où $P_V(D)$ est la probabilité d'observer ces données si le variant est pathogène, $P(V)$ est la probabilité *a priori* que le variant soit pathogène et $P(D)$ est la probabilité d'observer ces données dans la population générale. Pour s'affranchir du terme $P(D)$, dont l'estimation nécessiterait l'étude d'une grande cohorte, le rapport de vraisemblance est utilisé, appelé aussi *likelihood ratio* ou LR. La relation entre le LR et la probabilité *a posteriori* est définie par :

$$LR = \frac{P_D(V)}{1 - P_D(V)} = \frac{P_D(V)}{P_D(\bar{V})} \; et \; P_D(V) = \frac{LR}{LR + 1}$$

Où $\bar{V}$ correspond à l'évènement variant neutre. Ainsi le calcul du LR d'un variant, en prenant en compte les *i* données observées pour ce variant, correspond à :

$$LR = \frac{P(V)}{P(\bar{V})} \times \prod_i \frac{P_V(D_i)}{P_{\bar{V}}(D_i)}$$

Ainsi le calcul du LR ne dépend plus que de trois termes : la probabilité *a priori* que le variant soit pathogène ($P(V)$) et la probabilité d'observer ces données si le variant est neutre ou pathogène ($P_V(D_i)$ et $P_{\bar{V}}(D_i)$). La probabilité *a priori* est estimée le plus souvent à partir des prédictions bioinformatiques. Par exemple, l'outil AGVGD (*Align Grantham Variation and Grantham Deviation*) fourni un score de conservation de la séquence mutée [224], [225]. Ainsi une large cohorte de variants *BRCA1/BRCA2* a été utilisée pour estimer cette probabilité *a priori* [226]. Puis ces dernières ont été ensuite standardisées [227] (Figure 34).

Les probabilités $P_V(D_i)$ et $P_{\bar{V}}(D_i)$ sont estimées à partir de collectes de variants pathogènes et neutres associées avec leurs données cliniques et familiales [228]–[230]. Puis la probabilité *a posteriori* est calculée par un modèle multifactoriel comme décrit ci-dessus [231]. A noter que ce type d'approche est itérative, c'est-à-dire que la probabilité *a posteriori* d'une première analyse peut être utilisée comme une probabilité *a priori* pour une seconde analyse du même variant (Figure 34). Nous pouvons également remarquer que les variants situés dans les motifs canoniques d'épissage (notés SCS dans la Figure 34) ont une probabilité *a priori* d'être pathogène de 0.96. Cette probabilité indique que tout variant situé dans ces motifs canoniques est *a priori* probablement pathogènes (classe 4) indépendamment du modèle multifactoriel.

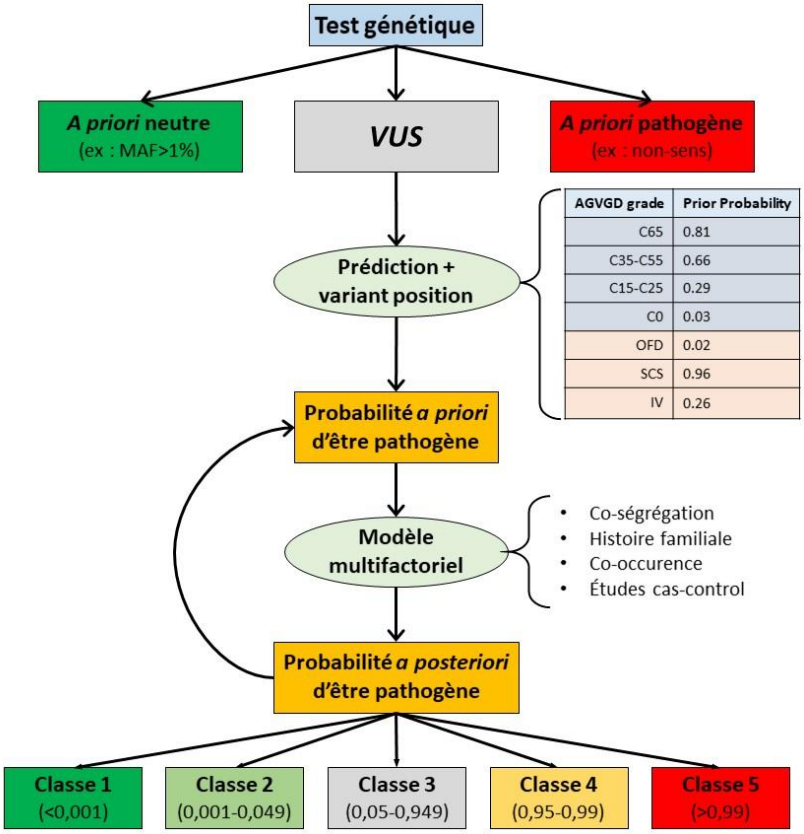


Figure 34 : Utilisation du modèle multifactoriel pour attribuer les 5 classes définies par le GCS (*Genetic Cancer Susceptibility*). Les probabilités *a priori* sont issues de [227]. C65-C0 : scores fournis par AGVGD (*Align Grantham Variation and Grantham Deviation*), OFD : *Outside functional domains*, SCS : *Splicing canonical site*, IV : *Intronic variants* (hors site canonique).

Cependant l'utilisation de modèle multifactoriel implique l'accès à un grand nombre de données pour chaque variant. De plus, l'estimation des probabilités $P(V)$, $P_V(D_i)$ et $P_{\overline{V}}(D_i)$ est étroitement liée aux choix des données utilisées. Ainsi pour homogénéiser l'attribution des classes de pathogénicité aux variants, notamment dans le cadre de maladies héréditaires rares, une méthode de classification catégorielle est également proposée. Elle suit les recommandations émise au niveau international par l'ACMG (*American College of Medical Genetics and Genomics*). Au niveau national l'ANPGM (*Association Nationale des Praticiens de Génétique Moléculaire*) a évalué et adapté ces recommandations par pathologie. L'ACMG propose de définir les principaux arguments utilisables pour classer les variants et d'établir le poids de ces arguments (Figure 35). En effet chacun d'entre eux est réparti dans les catégories suivantes en fonction de leur poids pour l'interprétation d'un variant : PM, *pathogenic moderate*, PP, *pathogenic supporting*, PS, *pathogenic strong*, PVS, *pathogenic very strong*. L'ACMG fournit également un algorithme décisionnel combinant ces arguments pour préciser la classe du variant (Figure 36).

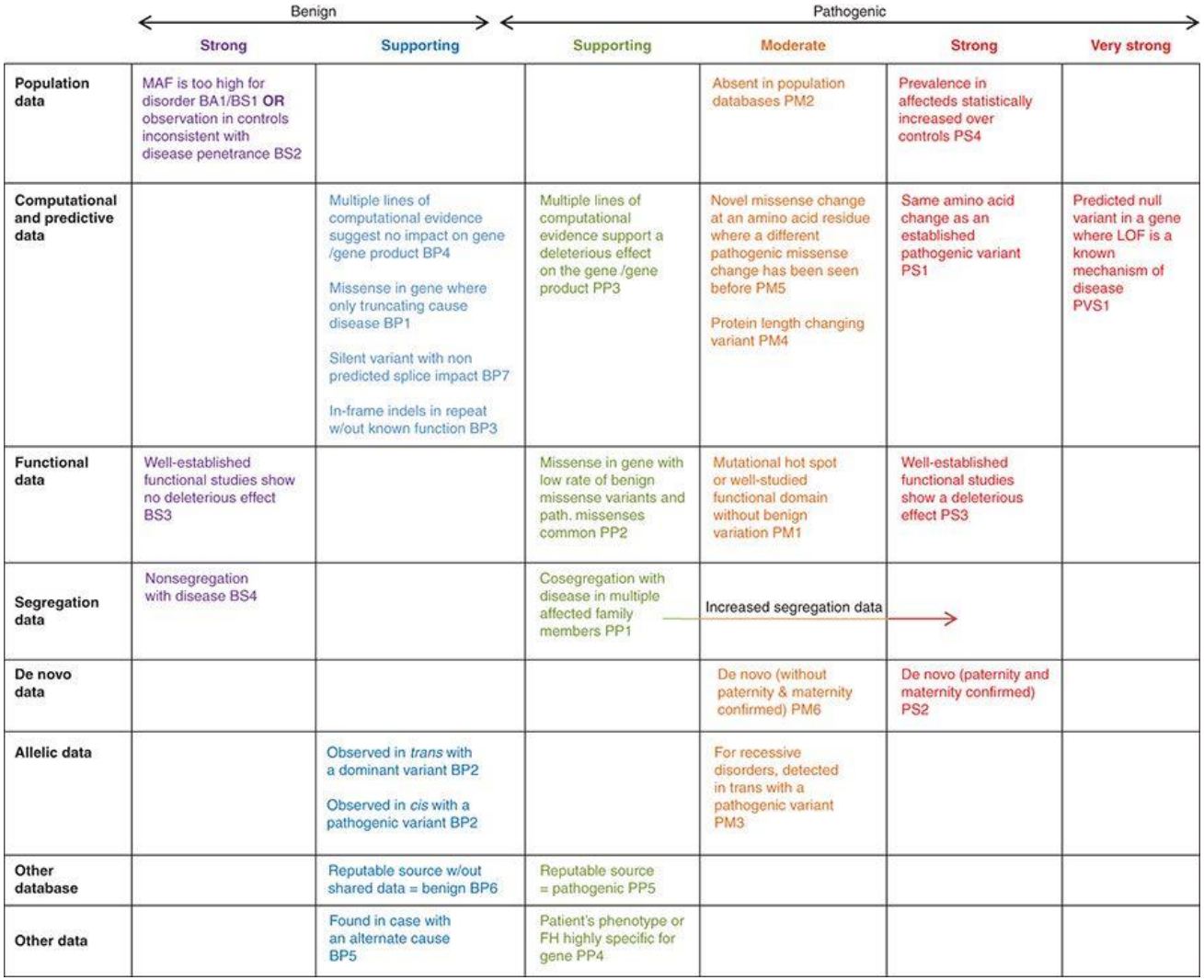


Figure 35 : Listes des principaux arguments utilisables ainsi que leur poids pour la classification des variants selon l'ACMG (*American College of Medical Genetics and Genomics*) (d'après [232]). BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

Pathogenic	(i) 1 Very strong (PVS1) *AND* (a) ≥1 Strong (PS1–PS4) *OR* (b) ≥2 Moderate (PM1–PM6) *OR* (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) *OR* (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) *OR* (iii) 1 Strong (PS1–PS4) *AND* (a)≥3 Moderate (PM1–PM6) *OR* (b)2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR* (c)1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Likely pathogenic	(i) 1 Very strong (PVS1) *AND* 1 moderate (PM1–PM6) *OR* (ii) 1 Strong (PS1–PS4) *AND* 1–2 moderate (PM1–PM6) *OR* (iii) 1 Strong (PS1–PS4) *AND* ≥2 supporting (PP1–PP5) *OR* (iv) ≥3 Moderate (PM1–PM6) *OR* (v) 2 Moderate (PM1–PM6) *AND* ≥2 supporting (PP1–PP5) *OR* (vi) 1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Benign	(i) 1 Stand-alone (BA1) *OR* (ii) ≥2 Strong (BS1–BS4)
Likely benign	(i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) *OR* (ii) ≥2 Supporting (BP1–BP7)
Uncertain significance	(i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory

Figure 36 : Algorithme décisionnel pour la classification des variants selon les arguments définis par l'ACMG (*American College of Medical Genetics and Genomics*) (d'après [232]).

Parmi les principaux arguments recommandés par l'ACMG, il y a les données de population. Ces dernières permettent de définir la fréquence allélique (MAF, *minor allele frequency*) du variant au sein de la population générale. Dans le cadre du syndrome HBOC, en dehors des effets fondateurs, un variant peut être considéré comme non pathogène si MAF > 0.1 %, d'après une réflexion initiée au sein du groupe génétique et cancer (GGC) (http://www.unicancer.fr/en/cancer-and-genetic-group). Parmi les variants entrainant un PTC dans *BRCA1/2*, la valeur maximale de la MAF est de 0,0277 % (*BRCA2* c.5946delT), données gnomAD (*Genome Aggregation Database*) (https://gnomad.broadinstitute.org/) [233]. Les données cliniques et notamment les données de ségrégation du variant au sein des familles sont exploitées pour attribuer la classe de pathogénicité au variant. Les prédictions *in silico* sont aussi prises en compte pour la classification des variants : prédiction de conservation nucléotidique, prédiction d'épissage et prédiction protéique. S'ils sont disponibles, les données de tests fonctionnels peuvent également être utilisées. Cependant à ce jour, il n'existe pas un unique test fonctionnel standard pour les

protéines BRCA1/BRCA2 étant donné la diversité de leurs rôles au sein de la cellule. En effet, il existe plusieurs tests fonctionnels pour mesurer l'activité de la protéine BRCA1 ou BRCA2 [234], [235] :

- Les tests fonctionnels basés sur l'utilisation de cellules souches embryonnaires de souris modifiées [236]–[238]. Ces cellules contiennent un allèle défectueux du gène d'intérêt et l'autre allèle fonctionnel mais pouvant être inactivé de façon conditionnelle. Le test repose sur l'observation qu'une protéine fonctionnelle est essentielle pour la survie de ces cellules. La transfection d'un plasmide contenant l'allèle portant la variation à tester permet d'observer ou non la survie cellulaire.
- Les tests basés sur l'activité de réparation directe par homologie [239], [240]. Ils reposent sur la présence de 2 allèles inactifs encodant le gène de la GFP dans le génome de cellules déficientes en *BRCA1* ou *BRCA2*. Un plasmide contenant l'allèle portant la variation à tester est transfecté dans la cellule, puis une cassure double-brin est induite au niveau de GFP par l'utilisation d'une endonucléase. Si la recombinaison homologue est efficace, la cellule exprimera alors la GFP.
- Les tests basés sur l'amplification du centrosome [241], [242]. En effet, l'inactivation des gènes *BRCA1* ou *BRCA2* provoque une amplification du nombre de centrosomes, visualisables par immunofluorescence.

Cependant, l'application de ces recommandations laisse une part non négligeable de variants en classe 3 (Figure 37). C'est-à-dire que pour les variants de cette classe, il n'est pas possible d'affirmer ou d'infirmer leur pathogénicité, ou que les modèles utilisés manquent de puissance. Ces variants nécessitent donc la réalisation d'analyses complémentaires telles que des tests fonctionnels (au niveau ARN et protéine) et/ou d'études cliniques complémentaires (collecte des données de co-ségrégation du variant avec la pathologie familiale). De nombreuses sociétés savantes sont impliquées dans la classification de ces VUS. Au niveau de l'oncogénétique, nous pouvons citer à l'échelle nationale le travail du GGC mené au sein de l'essai clinique COVAR (COségrégation VARiants) ainsi que le développement d'une base de données publique interprétative FROG (*French OncoGenetic database*). Pour chaque syndrome cette base sera alimentée par des données préexistantes de chaque groupe d'expert. Ainsi pour le syndrome HBOC, la base de données BRCA Share™ (http://www.umd.be/BRCA1/) [243] du GGC contribue à l'interprétation des variants. A l'échelle internationale le consortium ENIGMA (*Evidence-based Network for the Interpretation of Germline Mutant Alleles*) (https://enigmaconsortium.org/) [244] s'est spécialisé dans l'interprétation des variants en oncogénétique. Ce consortium permet notamment de mettre en relation différentes équipes, chacune partageant ses résultats afin de contribuer à classer ces VUS, de proposer des recommandations basées sur l'ACMG et de réévaluer les modèles multifactoriels [245]. Ceci a notamment permis de créer la base de données variants *BRCA1*/*BRCA2* BRCA exchange (https://brcaexchange.org/) [246].

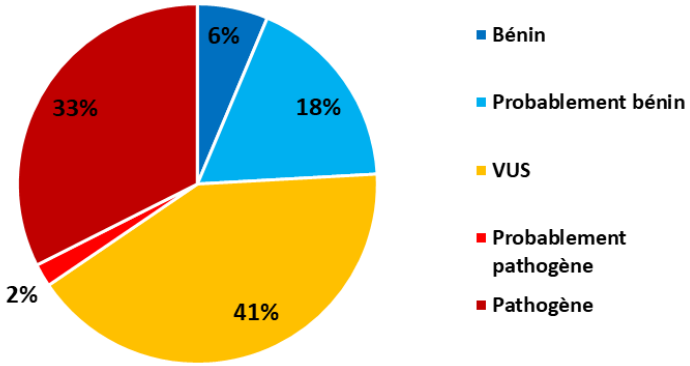


Figure 37 : Répartition des classes de pathogénicité pour les gènes *BRCA1*, *BRCA2*, *PALB2*, *RAD51C*/*D*. Données extraites de ClinVar, août 2019.

Malgré les recommandations internationales pour l'interprétation des variants, il n'est pas rare d'observer des discordances d'interprétation entre différents laboratoires [247]. A titre d'exemple, plus de 10 % des variants rapportés par au moins deux laboratoires ont une interprétation discordante dans la base de données ClinVar [248]. Ces données, ainsi que la proportion des VUS, illustrent que même actuellement l'interprétation clinique des variants reste un enjeu majeur.

3. Altération de l'épissage et pathogénicité : une histoire complexe

Comme nous l'avons précédemment décrit, les altérations de l'épissage par un variant sont diverses et nombreuses (voir section *Des variants génétiques aux défauts d'épissage*). Dans le cadre du syndrome HBOC, ces altérations de l'épissage sont pathogènes principalement par trois grands mécanismes : décalage du cadre de lecture avec génération d'un codon stop prématuré, perte d'une séquence codante pour un domaine fonctionnel et perte du codon d'initiation de la traduction. A titre d'exemple, le variant pathogène *BRCA1* c.547+2T>A est responsable du saut de l'exon 8 par abolition du site donneur canonique. Cet événement étant hors phase, un codon stop apparait 51 nt plus loin [249]. Le variant *BRCA1* c.5152+1G>T en détruisant le site donneur, occasionne un saut de l'exon 18, certes en phase, mais qui impacte le domaine fonctionnel majeur BRCT [250]. Ponctuellement, un défaut d'épissage peut être pathogène par insertion d'une séquence intronique en phase mais contenant un codon stop. Il peut également s'agir de variants modifiant les taux d'expression des épissages alternatifs avec un impact plus ou moins important sur la fonctionnalité de la protéine. Le variant c.316+2T>C du gène *BRCA2* produit un saut de l'exon 3 qui correspond à un renforcement majeur du transcrit alternatif non fonctionnel [191], [251], [252].

En regard de ces altérations d'épissage ayant un impact majeur sur la protéine, il est tout à fait possible d'observer des altérations de l'épissage ayant un impact mineur. Il peut s'agir, d'une part, des variants générant un saut d'exon en phase et ne comportant pas de domaine fonctionnel. Nous pouvons citer

l'exemple du variant c.6853A>G dans le gène *BRCA2*, qui par altération d'un ESR génère la perte de l'exon 12 décrit comme redondant [253]. Ainsi ce variant est classé neutre dans la base de données ClinVar. D'autre part, certains variants génèrent bien des sauts d'exon hors phase ou en phase, emportant un domaine fonctionnel, mais ceci de manière partielle. Cet effet partiel signifie que le transcrit muté est capable de générer à la fois un transcrit normal appelé transcrit pleine longueur (codant pour la protéine sauvage) et un transcrit anormal. A titre d'exemple le variant c.9501+3A>T dans le gène *BRCA2* ne produit qu'un saut partiel de l'exon 25 [95].

L'étape délicate de quantification de cet effet au cours de l'analyse de l'impact du variant est primordiale. Elle peut-être plus aisée si le gène étudié est porteur de polymorphismes ou d'autres variants exoniques ou si le variant étudié est exonique, car la présence ou l'absence des deux allèles peut alors être évaluée. Cependant, si le variant étudié est intronique, elle fait alors appel à l'analyse minigène. Des recommandations ont été émises par le GGC afin de standardiser les analyses et leur interprétation [254]. Cependant, la quantification de ces effets partiels ou totaux reste un défi technologique à relever afin d'améliorer l'interprétation de ces études ARN.

Enfin, un phénomène de compensation d'un effet potentiellement délétère, par un épissage alternatif (sans renforcement majeur de ce dernier) a récemment été mis en évidence [255]. Ainsi, le variant *BRCA1* c.594-2A>C, associé en *cis* avec le variant exonique c.641A>G sont à l'origine du saut hors phase de l'exon 10. Cet effet est total, donc potentiellement pathogénique. Cependant, ces variants ne co-ségrégent pas avec la pathologie. L'explication viendrait d'un phénomène de compensation par l'épissage alternatif correspondant au saut des exons 9 et 10. Cet évènement est présent à l'état physiologique à un taux d'environ 30% par rapport au transcrit pleine longueur. Ce transcrit semble fonctionnel, et serait à l'origine d'une absence d'haplo-insuffisance. Suite à cette étude, ce variant a été classé comme neutre dans la base de données ClinVar (revue par un panel d'expert). Ce mécanisme est illustré en Figure 38 A et B à titre d'exemple pour un transcrit pleine longueur contenant les exons 1 à 5 et un transcrit alternatif contenant les exons 1, 4 et 5. Ceci soulève également la question de la quantification du taux minimal de protéines fonctionnelles, traduites à partir du transcrit pleine longueur et des transcrits alternatifs pour une interprétation pertinente de la pathogénicité d'un variant altérant l'épissage.

Cette dualité des anomalies d'épissage implique également que les variants exoniques doivent d'abord être considérés pour leur impact potentiel au niveau de l'épissage avant d'évaluer l'impact protéique. En effet, un variant synonyme, ne modifiant pas théoriquement la séquence en acides aminés, peut avoir un impact sur l'épissage et ainsi être pathogène. A l'inverse, un variant non-sens, théoriquement délétère pour la protéine, peut ne pas l'être par la modification de l'épissage. Ce dernier cas correspond à un phénomène de sauvegarde de la protéine en utilisant l'épissage alternatif pour exclure l'exon porteur du PTC, lié au variant non-sens (Figure 38A et C). D'abord décrit sur le gène *DMD* de la dystrophie

musculaire de Duschenne [256], ce phénomène a tout récemment été démontré sur le gène *BRCA2* par renforcement du saut de l'exon 12 par l'équipe ARN Rouennaise de notre unité de recherche U1245, notamment pour le variant non-sens c.6901G>T (p.Glu2301X) qui altère potentiellement des SREs [257].

Ainsi, la connaissance des épissages alternatifs et de l'impact d'un variant sur l'épissage se révèle d'une importance majeure dans l'interprétation clinique de variants en génétique humaine.

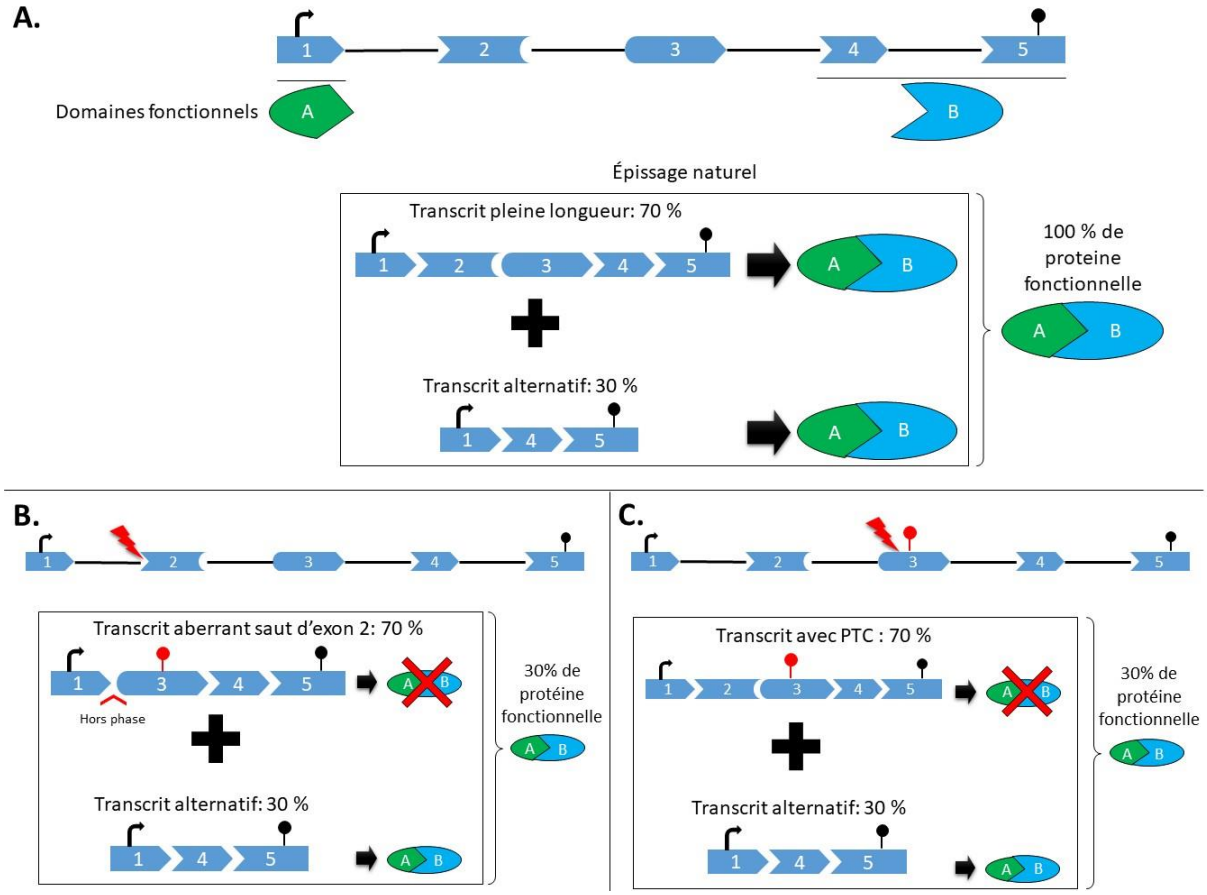


Figure 38 : Exemples de restaurations partielles de la fonctionnalité d'une protéine pour des variants théoriquement délétères par un épissage alternatif. **A.** Illustration d'un pré-ARNm avec 5 exons, l'exon 1 codant pour un premier domaine fonctionnel A et les exons 4 et 5 codants pour un second domaine fonctionnel B. Deux transcrits fonctionnels sont issus de ce pré-ARNm : un transcrit pleine longueur et un transcrit alternatif dépourvu des exons 2 et 3. **B.** Un variant sur le site accepteur de l'exon 2 est à l'origine d'une part de l'expression d'un transcrit anormal majoritaire (70%) comportant un saut hors phase de l'exon 2 et d'autre part de l'expression minoritaire (30%) du transcrit alternatif fonctionnel. **C.** Variant générant un PTC, le transcrit pleine longueur majoritaire (70%) portant le PTC est non fonctionnel tandis que le transcrit alternatif minoritaire (30 %) reste fonctionnel.

OBJECTIFS DES TRAVAUX DE THESE

De nos jours, l'utilisation des outils de NGS est entrée dans la routine clinique pour la recherche de variants nucléotidiques en génétique médicale. En effet, rien que dans les laboratoires français spécialisés en oncogénétique, plus de 90 % des recherches de variants sont actuellement réalisées par cette technologie [189]. Ainsi, la détection de nombreux évènements génétiques devient de plus en plus accessible. En contrepartie, le nombre de variants à interpréter par les biologistes croit de manière exponentielle. L'interprétation clinique de variants en génétique humaine peut avoir d'importantes conséquences sur la qualité de vie du patient [258]. Les défauts d'épissage imputables à un variant, sont une des principales sources d'interprétation inadéquate. En effet, les variants modifiant l'épissage sont sous-estimés pour plusieurs raisons notamment au niveau de leur nombre [46]. De plus, le volume conséquent de variants détectés par NGS et la difficulté d'accès à du matériel biologique rendent impossible l'étude ARN *in vitro* de façon systématique. Face à ce paradoxe, de nombreux outils de prédiction ont été proposés pour identifier les motifs d'épissage et prédire leurs altérations par un variant.

Cependant, actuellement, les biologistes sont confrontés à une grande diversité d'outils de prédiction chacun intégrant différentes méthodes et étant spécifique d'un ou plusieurs motifs d'épissage. De sorte que l'intégration de ces outils dans le processus d'interprétation des variants requière des recommandations. Ces dernières peuvent ainsi proposer un outil ou une combinaison d'outils optimaux pour chaque motif d'épissage et des seuils de scores décisionnels afin de retenir ou non un variant pour une étude ARN *in vitro*. Les premières recommandations pour l'utilisation des outils de prédiction d'épissage pour les motifs consensus donneur/accepteur datent de 2007. Mais, elles sont basées sur un nombre restreint de variants et/ou spécifiques d'un gène ou d'un panel de gènes associés au même syndrome [53], [259].

En ce qui concerne le syndrome HBOC, le réseau épissage du GGC collecte des études ARN *in vitro*, au sein duquel notre laboratoire participe activement. Grâce à ce travail, en 2012, il a été publié la plus grande étude de collecte de ce type de variants. Elle a permis de dégager des recommandations sur l'utilisation des outils de prédiction d'épissage pour les motifs consensus donneur/accepteur ainsi que pour la réalisation et l'interprétation des études ARN. Celle-ci s'est basée sur l'étude ARN de près de 300 variants *BRCA1/BRCA2*, dont 65 situés dans ces motifs consensus [254]. Il en a résulté que les outils optimaux étaient MES et SSF. Les seuils décisionnels étaient de 15 % pour MES et de 5 % pour SSF de diminution de score entre les séquences sauvages et mutées. Depuis, le réseau épissage du GGC a continué à collecter des données ARN expérimentales afin de pouvoir réévaluer ces recommandations. Ainsi, un des premiers objectifs de cette thèse a été de réévaluer les outils de prédiction dédiés aux sites consensus donneur/accepteur à partir de cette nouvelle collection. De plus, nous nous sommes aussi associés au consortium international ENIGMA ainsi qu'à d'autres équipes de recherche dans le but

d'augmenter la puissance du jeu de données et de pouvoir tester ces résultats sur des gènes impliqués dans des pathologies humaines autres que le syndrome HBOC. Nous avons ainsi pu réexaminer ces recommandations sur 395 variants dans les motifs consensus donneur/accepteur contre 65 variants en 2012. Nous avons ainsi développé un outil simple d'utilisation permettant de prédire l'impact sur l'épissage afin de prioriser les études ARN *in vitro*, appelé SPiCE pour *Splicing Prediction in Consensus Elements* [260].

Cependant, si nous disposons de recommandations sur l'utilisation des outils de prédiction pour les motifs consensus d'épissage donneur/accepteur, il n'en est pas de même pour les autres signaux d'épissage. La prédiction des sites de branchement illustre parfaitement cette problématique. Le fait que les motifs des sites de branchement soient de courtes séquences dégénérées et de positions variables, rend leur détection et la prédiction de leurs altérations particulièrement complexes. De plus, ce n'est qu'à partir de 2015 que nous avons disposé d'une collection importante de points de branchement validés expérimentalement [15]. Cette collection a ainsi récemment motivé la genèse de plusieurs outils de prédiction dédiés au point de branchement [166]–[168], [261]. Aussi, au cours de cette thèse, nous nous sommes proposés d'évaluer ces outils pour leur capacité à détecter la présence de point de branchement et aussi pour leur capacité à prédire l'altération de ces points de branchement par un variant. Pour cela, nous avons utilisé des données issues de transcrits connus, de données de RNA-seq ainsi que d'une collection de plus de 100 variants avec leur données ARN expérimentales (Leman *et al*, article en cours de révision, BMC Genomics) [262].

Dès lors que les variants sont sélectionnés pour leur potentiel impact sur l'ARN, se pose la question de la méthodologie pour caractériser le défaut d'épissage imputable à ces variants. Si encore aujourd'hui la RT-PCR reste une méthode largement utilisée, le RNA-seq offre la possibilité de détecter un plus large éventail d'évènement d'épissage. En particulier le RNA-seq ciblé se révèle notoirement intéressant pour identifier la diversité des transcrits issus des gènes d'intérêt. Il a ainsi été utilisé avec succès au sein de notre laboratoire pour la description fine des épissages alternatifs de 11 gènes impliqués ou suspectés être impliqués dans le syndrome HBOC [131].

Les données générées par RNA-seq sont plus complexes et volumineuses que celles générées par RT-PCR. Aussi, l'automatisation des analyses des jonctions d'épissage vers des résultats humainement exploitables est indispensable pour une possible intégration du RNA-seq dans le processus d'interprétation clinique des variants. De nombreux outils ont déjà été développés pour quantifier et détecter efficacement une différence d'expression des gènes (revue par [140]). Toutefois, pour les jonctions d'épissage, l'analyse se révèle plus complexe car elle doit considérer des évènements chevauchants et le plus souvent non décrits dans de précédentes base de données [263]. Ainsi, dans le cadre de cette thèse nous avons développé un outil bioinformatique et biostatistique pour l'annotation,

la quantification et la détection de jonctions anormales d'épissage (Leman *et al*, article en cours de révision, Bioinformatics).

Suite à l'identification des évènements d'épissage, se pose la question de leur importance lors de l'interprétation clinique des variants. Un variant situé dans un motif canonique d'épissage (ex : AG/GT), reconnu comme modifiant nécessairement l'épissage, est *a priori* considéré comme probablement pathogène [227]. Cependant, cette démarche ne prend pas en compte l'existence des épissages alternatifs de chaque gène. Or un variant peut les renforcer au lieu de générer un transcrit aberrant. Sachant que ces épissages alternatifs peuvent être fonctionnels, cela soulève la question de la réelle pathogénicité d'un tel variant. Le gène *PALB2* illustre bien cette problématique. Il a récemment été introduit en diagnostic pour son association au syndrome HBOC pour lequel à ce jour nous ne disposons que de peu de données ARN. Ainsi, un projet collaboratif a été mis en place par le consortium ENIGMA. Ce projet portait sur deux aspects : l'identification des épissages alternatifs de ce gène et l'étude exhaustive de l'impact des variants canoniques sur ces épissages alternatifs au regard des critères d'interprétation de l'ACMG. Au cours de ce travail de thèse, nous avons générés les données d'épissages alternatifs identifiés par RNA-seq dans 72 échantillons. Ces échantillons provenaient de sang total, de leucocytes et de tissus mammaires. Le premier objectif était d'évaluer d'une part si l'expression de *PALB2* est tissu spécifique. Le second était d'estimer si un variant PTC peut par un phénomène de compensation ne pas être pathogène en raison de la présence d'épissages alternatifs. Le troisième était d'apprécier si des variants situés au niveau des régions canoniques (AG, GT) pouvaient induire la production de transcrits alternatifs en phase et fonctionnels. Ce travail a permis la publication d'une cartographie précise des épissages alternatifs du gène *PALB2* et des recommandations concernant l'interprétation de l'impact potentiel de variants PTC et d'épissage en absence d'étude de l'ARN [264].

Par conséquent au cours de ce travail de thèse, nous nous sommes intéressé à trois grands aspect de l'étude des défauts de l'épissage : la prédiction de ces défauts d'épissage, l'analyse des données de RNA-seq et l'implication des transcrits alternatifs dans l'interprétation de la pathogénicité d'un variant.

RESULTATS

Dans cette partie nous allons détailler les résultats des travaux de cette thèse. Ces derniers ont conduit à la publication et soumission de 5 articles (3 publiés, 1 en cours de révision, 1 en cours de rédaction) :

I. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined *in silico/in vitro* studies: an international collaborative effort. *(publié)*

II. Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants. *(en révision)*

III. SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations, validated on a curated diagnostic set of 2,784 exonic and intronic variants. *(en cours de rédaction)*

IV. SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data. *(publié)*

V. Alternative Splicing and ACMG-AMP-2015 Based Classification of *PALB2* Genetic Variants: an ENIGMA Report. *(publié)*

I. Nouvel outil diagnostique pour la prédiction de variants splicéogéniques situés dans les sites consensus : Article I

Le présent travail a été publié dans *Nucleic Acids Research* en 2018 (doi : https://doi.org/10.1093/nar/gky979). Les données supplémentaires de cet article sont en **ANNEXE A**, les tableaux contenant la liste des variants avec leurs prédictions et leurs données ARN (*tables S1 to S3*) sont en ligne à https://academic.oup.com/nar/article/46/21/11656/5128933#supplementary-data.

L'interprétation des variants est un enjeu majeur pour le diagnostic moléculaire. Cependant, les généticiens sont confrontés à un nombre grandissant de VUS pour lesquels il n'est pas possible de conclure sur la pathogénicité. Les variants splicéogéniques illustrent cette problématique car chaque variant peut-être pathogène par la création ou la destruction de motif d'épissage. L'altération des motifs consensus donneur/accepteur est une des principales causes d'anomalie d'épissage. De nombreux outils de prédiction ont donc été développés pour prédire cette altération. En 2012 le réseau épissage du GGC avait publié des premières recommandations pour l'utilisation de ces outils pour les gènes *BRCA1* et *BRCA2* [254]. Depuis 2012, de nombreux variants ont été collectés ainsi que les données de leur étude ARN *in vitro*. L'application de ces recommandations a généré un taux de faux négatifs conséquent (11,8%), amenant à les reconsidérer pour une application diagnostique.

Grâce à un effort collaboratif international, 395 variants ont été collectés avec leurs études ARN *in vitro* dans les sites consensus de 11 gènes (incluant *BRCA1*, *BRCA2*, *CFTR* et *RHD*). Cette collection nous a permis de comparer et de tester différentes combinaisons de scores de prédiction dédiés aux sites d'épissage consensus. Nous avons développé un nouvel outil pour la prédiction d'un défaut de l'épissage par un variant indépendamment du gène, nommé SPiCE (*Splicing Prediction in Consensus Element*). Ainsi, ces variants ont été divisés en deux jeux de données. Un jeu de données, comprenant les variants des gènes *BRCA1/BRCA2* collectés par le réseau épissage du GGC (n = 142), a été utilisé pour l'apprentissage de SPiCE. Le second jeu de données comprenant à la fois des variants *BRCA1/BRCA2* et non *BRCA1/BRCA2* (n = 253), a été utilisé pour la validation de SPiCE. Les scores de prédiction considérés pour ce travail étaient SSF, NNS, GS, MES et HSF.

Ainsi nous avons pu montrer que l'intégration des scores MES et SSF par régression logistique dans SPiCE permet d'améliorer significativement la prédiction des défauts d'épissage. En effet, SPiCE atteint sur le jeu de validation, une exactitude de 98.8 %. Nous avons également défini deux seuils décisionnels optimaux. Le premier a été défini dans un but d'application diagnostique avec une sensibilité optimale (99.5 %). Le second s'adresse à un contexte de recherche biomédicale avec une spécificité optimale (95.2 %). Ainsi SPiCE a permis de corriger le nombre de faux négatifs générés par les précédentes recommandations. SPiCE est accessible sous la forme d'un logiciel interfacé accessible à https://sourceforge.net/projects/spicev2-1/.

Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined *in silico*/*in vitro* studies: an international collaborative effort.

Raphaël Leman[1,2,3+#], Pascaline Gaildrat[2+#], Gérald Le Gac[4], Chandran Ka[4], Yann Fichou[4], Marie-Pierre Audrezet[4], Virginie Caux-Moncoutier[5,6+], Sandrine M. Caputo[7+#], Nadia Boutry-Kryza[8+], Mélanie Léone[8+], Sylvie Mazoyer[9+#], Françoise Bonnet-Dorion[10+], Nicolas Sevenet[10+], Marine Guillaud-Bataille[11+], Etienne Rouleau[11+], Brigitte Bressac-de Paillerets[11+], Barbara Wappenschmidt[12#], Maria Rossing[13#], Danielle Muller[14+], Violaine Bourdon[15+], Françoise Revillon[16+], Michael T. Parsons[17#], Antoine Rousselin[1,2+], Grégoire Davy[1,2+], Gaia Castelain[2+], Laurent Castéra[1,2+], Joanna Sokolowska[18+], Florence Coulet[19+], Capucine Delnatte[20+], Claude Férec[4], Amanda B Spurdle[17#], Alexandra Martins[2+#], Sophie Krieger[†1,2,3+#*], Claude Houdayer[†5,6,7+#*]

[†]These authors contributed equally to this work.

Unicancer Genetic Group (UGG)[+] splice network and ENIGMA[#]

[1]Laboratoire de Biologie Clinique et Oncologique, Centre François Baclesse, France, [2]Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders, UNIROUEN Normandie Université, Normandy Centre for Genomic and Personalized Medicine, Rouen, France, [3]Université Caen-Normandie, France, [4]Inserm UMR1078, Genetics, Functional Genomics and Biotechnology, Université de Bretagne Occidentale, Brest, France, [5]Inserm U830, Centre de Recherches, Paris, France, [6]Université Paris Descartes, Sorbonne Paris Cité, Paris, France, [7]Service de Génétique, Institut Curie, Paris, France, [8]Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon, France, [9]Lyon Neuroscience Research Center–CRNL, Inserm U1028, CNRS UMR 5292, University of Lyon, Lyon, France, [10]Inserm U916, Département de Pathologie, Laboratoire de Génétique Constitutionnelle, Institut Bergonié, Bordeaux, France, [11]Gustave Roussy, Université Paris-Saclay, Département de Biopathologie, Villejuif, France, [12]Division of Molecular Gynaeco-Oncology, Department of Gynaecology and Obstetrics, University Hospital of Cologne, Cologne, Germany, [13]Centre for Genomic Medicine, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark, [14]Laboratoire d'Oncogénétique, Centre Paul Strauss, Strasbourg, France, [15]Laboratoire d'Oncogénétique Moléculaire, Institut Paoli-Calmettes, Marseilles, France, [16]Laboratoire d'Oncogénétique Moléculaire Humaine, Centre Oscar Lambret, Lille, France, [17]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Herston, Queensland, Australia, [18]Service de Génétique, CHU Nancy, Nancy, France, [19]Service de génétique, Hôpital Pitié Salpétrière, AP-HP, Paris, France, [20]Laboratoire de génétique moléculaire, CHU Nantes, Nantes, France

*To whom correspondence should be adressed. Email: claude.houdayer@curie.fr and S.KRIEGER@baclesse.unicancer.fr

Present address: Sophie Krieger, Laboratoire de biologie et génétique des cancers, Centre François Baclesse, Caen, France ; Claude Houdayer, Service de Génétique, Institut Curie, Paris, France

1. ABSTRACT

Variant interpretation is the key issue in molecular diagnosis. Spliceogenic variants exemplify this issue as each nucleotide variant can be deleterious *via* disruption or creation of splice site consensus sequences. Consequently, reliable *in silico* prediction of variant spliceogenicity would be a major improvement. Thanks to an international effort, a set of 395 variants studied at the mRNA level and occurring in 5' and 3' consensus regions (defined as the 11 and 14 bases surrounding the exon/intron junction, respectively) was collected for 11 different genes, including *BRCA1*, *BRCA2* and *CFTR*, and *RHD* used to train and validate a new prediction protocol named Splicing Prediction in Consensus

Elements (SPiCE). SPiCE combines *in silico* predictions from SpliceSiteFinder-like and MaxEntScan and uses logistic regression to define optimal decision thresholds. It revealed an unprecedented sensitivity and specificity of 99.5% and 95.2%, respectively and the impact on splicing was correctly predicted for 98.8% of variants. We therefore propose SPiCE as the new tool for predicting variant spliceogenicity. It could be easily implemented in any diagnostic laboratory as a routine decision making tool to help geneticists to face the deluge of variants in the next-generation sequencing era. SPiCE is accessible at **(**https://sourceforge.net/projects/spicev2-1/**).**

2. INTRODUCTION

Since the advent of genome wide sequencing, interpretation of variants of unknown significance has been recognized as the major bottleneck and challenge for clinical geneticists. Variants are usually classed within a 5-tiered scheme [223] from benign and likely benign variants (class 1 and 2, respectively) to likely pathogenic and pathogenic variants (class 4 and 5, respectively). The geneticist is on relatively solid ground in these 4 classes, where the biological impact is known or at least likely known. However, class 3 refers to the so called Variants of Unknown Significance (VUS) where the effect of the sequence variation on the transcript and protein and thereby on the patient is simply not known. Clinical management logically stems from this knowledge [232] which is why variant classification is of utmost importance.

Hereditary breast and ovarian cancers are mainly due to *BRCA1* (MIM #113705) and *BRCA2* (MIM #600185) pathogenic variants. The *BRCA* genes embody the problem of variant interpretation due to their wide mutational spectrum, which is mostly devoid of specific hot spots. To exemplify this issue, over 30% of the variants in the Breast Cancer Information Core (BIC), ClinVar and BRCA Share databases are VUS [144], [265], [266].

Spliceogenic variants are probably the most challenging for the geneticists as each nucleotide variation, regardless of its location, can potentially affect pre-ARNm splicing and be pathogenic via disruption of 5’ or 3’ splice sites (5’/3’ ss), creation of new 5’/3’ ss or alteration of splicing regulatory elements. It is estimated that ~15% of all point mutations causing human inherited disorders disrupt splice-site consensus sequences [267]. Consequently, assessing the impact of variants on splicing is a mandatory task in molecular diagnosis. Towards this aim, several *in silico* prediction tools can be used either as stand-alone programs or as interfaces integrating multiple algorithms (see Materials and Methods). These tools are important to select variants that are worthy of expensive and time-consuming RNA analyses. This is why we published user’s guidelines from the splice network of French *BRCA* diagnostic laboratories within the Unicancer Genetic Group hereinafter named UGG, http://www.unicancer.fr/en/unicancer-group) [254], recommending the combined use of two bioinformatics variation scores MaxEntScan (MES) and Splice Site Finder-like (SSF-like) between the mutated and wild type sequences. Two thresholds of relative decrease of scores at 15% for MES and

5% for SSF-like permitted to obtain a sensitivity of 96% and a specificity of 83%. While useful, these guidelines are prone to false-negative predictions (see below, "results" section) and could therefore be improved. Consequently, we developed a new prediction tool, called Splicing Predictions in Consensus Elements (SPiCE), to prioritize RNA studies to relevant variants that alter 5' and 3' splice consensus regions *i.e* 11 bases for the 5' splice site and 14 bases for the 3' splice site. SPiCE uses logistic regression by running different combinations of *in silico* tools. Thanks to an international collaborative effort including the ENIGMA consortium (Evidence-based Network for the Interpretation of Germline Mutant Alleles, https://enigmaconsortium.org/) [244], we were able to collect 305 *BRCA1* and *BRCA2* variants occurring in 5' and 3' consensus regions with their corresponding splice study. SPiCE was developed using a training set of 142 *BRCA1* and *BRCA2* variants and validated on a further set of 163 *BRCA1* and *BRCA2* splice variants. Furthermore, and to demonstrate its versatility, SPiCE was successfully applied to another set of 90 variants occurring in 5' and 3' consensus regions of 9 non cancer genes *e.g*. in *CFTR* (MIM#602421), *CTRC* (MIM#601405), *HFE* (MIM#613609), *HJV* (MIM#608374), *LRP5* (MIM#603506), *PDK1* (MIM#602524), *RHD* (MIM#111690), *SLC40A1* (MIM#604653) and *TFR2* (MIM#604250).

3. MATERIALS AND METHODS

a. Nomenclature

Nucleotide numbering is based on the cDNA sequence of *BRCA1, BRCA2, CFTR, CTRC, HFE, HJV, LRP5, PKD1, RHD, SLC40A1, TFR2* (NCBI accession number NM_007294.2, NM_000059.3, NM_000492.3, NM_007272.2, NM_000410.3, NM_213653.3, NM_002335.3, NM_001009944.2, NM_016124.4, NM_014585.5, NM_003227.3, respectively), c.1 denoting the first nucleotide of the translation initiation codon, as recommended by the Human Genome Variation Society.

b. Definition of consensus splice site regions

Consensus splice site regions (5'ss and 3'ss) were defined according to Burge *et al*., 1999 [13], *i.e.* 11 bases for the 5' splice site (from the 3 last exonic to the 8 first intronic bases) and 14 bases for the 3' splice site (from the 12 last intronic to the first 2 exonic bases).

c. Datasets

Among this initiative, 395 variants occurring in the consensus 5'/3' ss regions of 11 genes were collected, along with their respective RNA studies, and distributed between a training set and a validation set (Figure 39).

The training set (Table S1) comprises 142 *BRCA1* and *BRCA2* variants from the UGG network. We performed transcript analyses as previously described [254]. Briefly, protocols for transcript analyses included i) minigene-based splicing assays, ii) RNA extracted from lymphoblastoid cell lines

treated/untreated with puromycin. iii) RNA extracted from blood collected into PAXgene tubes (Qiagen), iv) RNA extracted from stimulated T lymphocytes. Controls (samples without variant) were always included in these experiments. No discordance was observed between in vitro studies for the same variants.

To validate the SPiCE tool, we first gathered from the literature 208 transcript analyses from 163 distinct *BRCA1* (n = 92) and *BRCA2* (n = 71) variants reported in 56 publications. This curated collection of information was provided by members of the ENIGMA consortium as part of an ongoing data collection used for variant review [268], [269] (Table S2). Twelve of them (denoted by cross (†) in Table S2) were analyzed at least twice and splicing alteration was constantly observed for 11 variants, with outcomes for different variants including exon skipping, use of cryptic splice site or combination of these events. Only one variant (c.518G>T in *BRCA2*) had contradictory reports and the reasons for this discordance remain unknown [150], [270]. Secondly, to extend the use of SPiCE to non *BRCA*-genes, the second set of validation comprised 90 variants on *CFTR* (n = 44), *CTRC* (n = 2), *HFE* (n = 1), *HJV* (n = 1), *LRP5* (n = 1), *PKD1* (n = 1), *RHD* (n = 38), *SLC40A1* (n = 1) and *TFR2* (n = 1) with their splicing effect evaluated by minigene assay (Table S3) [271]. These variants were identified during the course of genetic counseling and thereby reflect clinical practice.

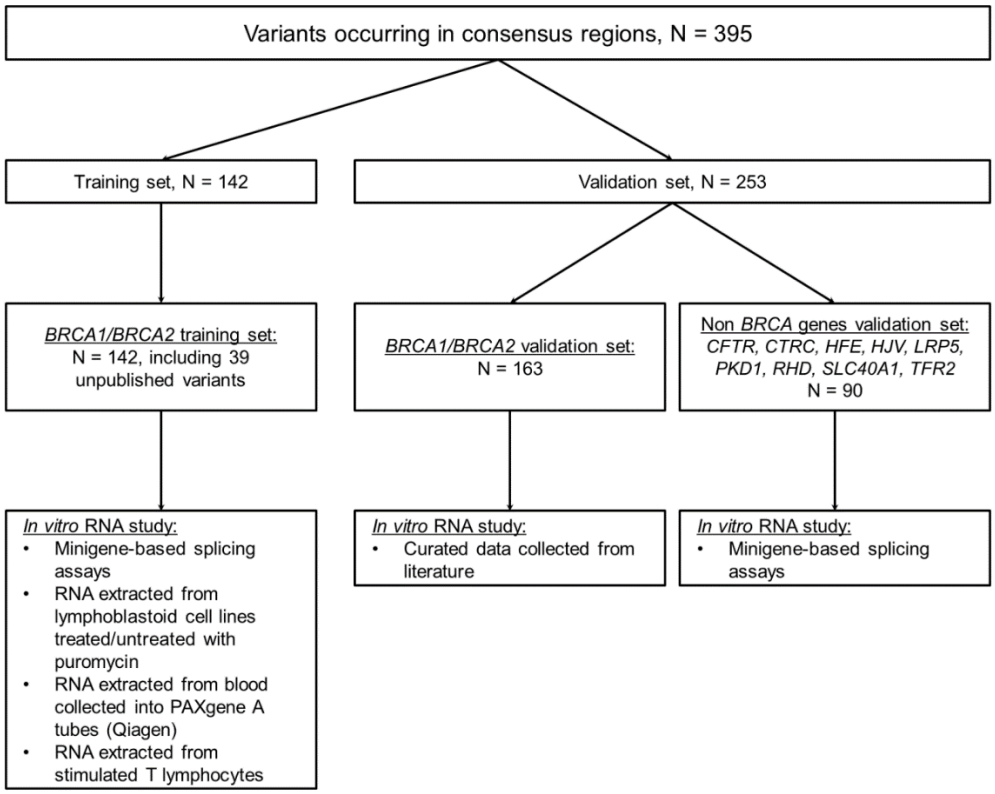


Figure 39 : Curated datasets and *in vitro* analyses methods used in this study

d. In silico tools

Five *in silico* prediction tools were tested: MaxEntScan (MES) , (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html) [148], Splice Site Finder (SSF) [145], Human Splicing Finder (HSF) (http://www.umd.be/HSF3/) [146], Neural Network Splice (NNS) (http://www.fruitfly.org/seq_tools/splice.html) [162] and GeneSplicer (GS) (http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml) [155]. Very briefly, the calculation of a MES score is based on maximum entropy of a nucleotide sequence with a set of constraints fixed by the MES model, including the variant's neighboring bases. NNS also takes into account the variant's neighboring position, but unlike MES, NNS is based on a machine learning technique *i.e.* artificial neural networks. For SSF and HSF, the score calculation is based on a position weight matrix and its homologous percentage with the tested sequence. We used SSF-like, a version of SSF, allowing calculation score of donor splice site with GT and GC canonical motifs, embedded in Alamut® and in SPiCE. Finally, GS is based on a decision tree method. It captures potential strong dependencies between signal positions by dividing the dataset into subsets based on pairwise dependency between positions and modeling each subset separately [178]. The outcomes of each of these tools were simultaneously obtained by using the commercial software (Alamut® Visual software version 2.8 rev. 1 and Alamut® Batch version 1.5.2., Interactive Biosoftware).

e. Logistic regression and model definition

First, we processed to descriptive analysis of bioinformatic prediction scores. We tested the discriminant capacity of these scores by Receiver-Operating Characteristics (ROC) curves, representing the sensitivity as a function of 1-specificity [176], using the R package ROCR [272], and the correlation between variables by Pearson's coefficient. Then, we used logistic regression to estimate the probability that a variant alters splicing. Parameters values were obtained by maximum likelihood, as objective function. This model was implemented in R software version 3.3.1 with the generalized linear model (glm) function. We considered that splicing alterations could correspond either to abnormal splicing events or to reinforcement of alternative splicing with partial or total effect. Splice event can be a single or multiple exon skipping and the use of exonic or intronic cryptic 5' or 3' splice sites. Selected variables to explain splicing alteration by a variant were i) variation of prediction scores between wild type (WT) and variant sequences, defined by Equation (1) and the score was annotated ΔMES, ΔSSF, ΔHSF, ΔNNS, or ΔGS, ii) localization in the invariant splice site positions (3'AG/5'GT), iii) donor (5') or acceptor (3') splice sites, iv) genes (*e.g.*: *BRCA1* or *BRCA2*).

$$\Delta score = \frac{score_{mutated} - score_{wt}}{score_{wt}} \quad (1)$$

To construct our final model we used a selection procedure based on a stepwise type approach with Akaike Index Criterion (AIC). Thereby AIC allows us to consider the likelihood of our model and the

number of parameters in order to have the best model with a minimum of parameters. Models were compared by a likelihood ratio test. Cross-validation and other validation steps of the final model are described in the supplementary methods. AIC was considered more relevant than the Bayesian Information Criterion for a predictive approach. In any case the two different criteria provided similar values (see table S7).

We developed SPiCE software, in the commonly utilized 'R' language to enable it to be freely applicable, information on this software are in supplementary material (see SPiCE handbook supplementary document). This software generates MES and SSF-like scores. For this purpose, the MES script was retrieved from the BurgeLab website (see *In silico* tools) and the SSF-like script was rewritten for SPiCE in R language according to its description under the original publication [145] and under the manual of the commercial Alamut software. Position weight matrices, used by SSF-like for scoring acceptor and donor splice sites, were obtained from SpliceDB which contains 28 468 pairs of splice site sequences [273].

f. In silico predictions using previously published guidelines

In order to compare SPiCE with our former guidelines, the *BRCA1/2* validation set was assayed as previously described [254].

4. RESULTS

Aberrant splicing events were described for each dataset in table S4. Briefly we observed 76.7% (303/395) variants that alter splicing, with 44.6% of exon skipping, 10.9% use of 5' alternative splice sites, 8.9% use of 3' alternative splice site and 12.4% of multiple aberration.

a. BRCA1/BRCA2 training set

In total we performed 188 *in vitro* analyses on 142 variants including 37 unpublished variants on both *BRCA1* (21 variants) and *BRCA2* (16 variants). The variants from the training set were equally distributed between *BRCA1* and *BRCA2* genes, 50.7% (72/142) and 49.3% (70/142) respectively. Eighty-four variants (60%) were localized at the proximity of the 5' ss and the 58 (40%) remainder at the proximity of the 3' ss. Ninety-five variants altered splicing and were mainly (54.7%, 52/95) located outside the AG/GT dinucleotides (Table 1 and Figure 40A).

b. BRCA1/BRCA2 validation set

In the 163 variants collected from the literature, 92 (56.4%) variants were in *BRCA1* and the 71 variants in *BRCA2*. These variants were mainly localized on the donor sites compared to the acceptor sites, 58.3% (94/163) and 41.7% (69/163), respectively. Sixty of 135 (44.4%) variants that alter splicing were outside canonical dinucleotides (Table 1 and Figure 40B).

c. Non-BRCA validation set

We also selected 90 variants in nine non-*BRCA* genes, which were in *CFTR* (n = 44), *CTRC* (n = 2), *HFE* (n = 1), *HJV* (n = 1), *LRP5* (n = 1), *PKD1* (n = 1), *RHD* (n = 38), *SLC40A1* (n = 1) and *TFR2* (n = 1) (Table S3). Fifty-three variants (58.9%) were in donor splice sites and 37 (41.1%) in acceptor sites. Seventy-three variants altered splicing in minigene assays. Half of these (n =36; 49.3%) are in the AG/GT dinucleotides (Table 1 and Figure 40C). Some positions were poorly represented and this uneven distribution outside 5'/3' ss can explain the imbalance between variants that do and do not affect splicing, 73 and 17 variants respectively (Figure 40C).

Table 1 : Distribution of variants in training and validation sets (n = 395)

	No. (%) of variants			**No. (%) of variants altering splicing**	
	5'/3' splice site			5'/3' splice site	
Gene	5'	3'	Gene	5'	3'
***Training set,* n = 142 variants**			**n = 95 variants altering splicing**		
BRCA1	42 (58.3)	30 (41.7)	*BRCA1*	32 (66.7)	16 (33.3)
BRCA2	42 (60.0)	28 (40.0)	*BRCA2*	32 (68.1)	15 (31.9)
Total	84 (59.2)	58 (40.8)	Total	64 (67.4)	31 (32.6)
***BRCA1/BRCA2 validation set,* n = 163**			**n = 135 variants altering splicing**		
BRCA1	54 (58.7)	38 (41.3)	*BRCA1*	49 (64.5)	27 (35.5)
BRCA2	40 (56.3)	31 (43.7)	*BRCA2*	36 (61.0)	23 (39.0)
Total	94 (57.7)	69 (42.3)	Total	85 (63.0)	50 (37.0)
***Non BRCA validation set,* n = 90**			**n = 73 variants altering splicing**		
CFTR	23 (52.3)	21 (47.7)	*CFTR*	23 (60.5)	15 (39.5)
RHD	26 (68.4)	12 (31.6)	*RHD*	22 (73.3)	8 (26.7)
Other genes[a]	4 (50.0)	4 (50.0)	Other genes[a]	3 (60.0)	2 (40.0)
Total	53 (58.9)	37 (41.1)	Total	48 (65.8)	25 (34.2)

[a] : *LRP5, CTRC, HFE, HJV, PKD1, SLC40A1, TFR2*

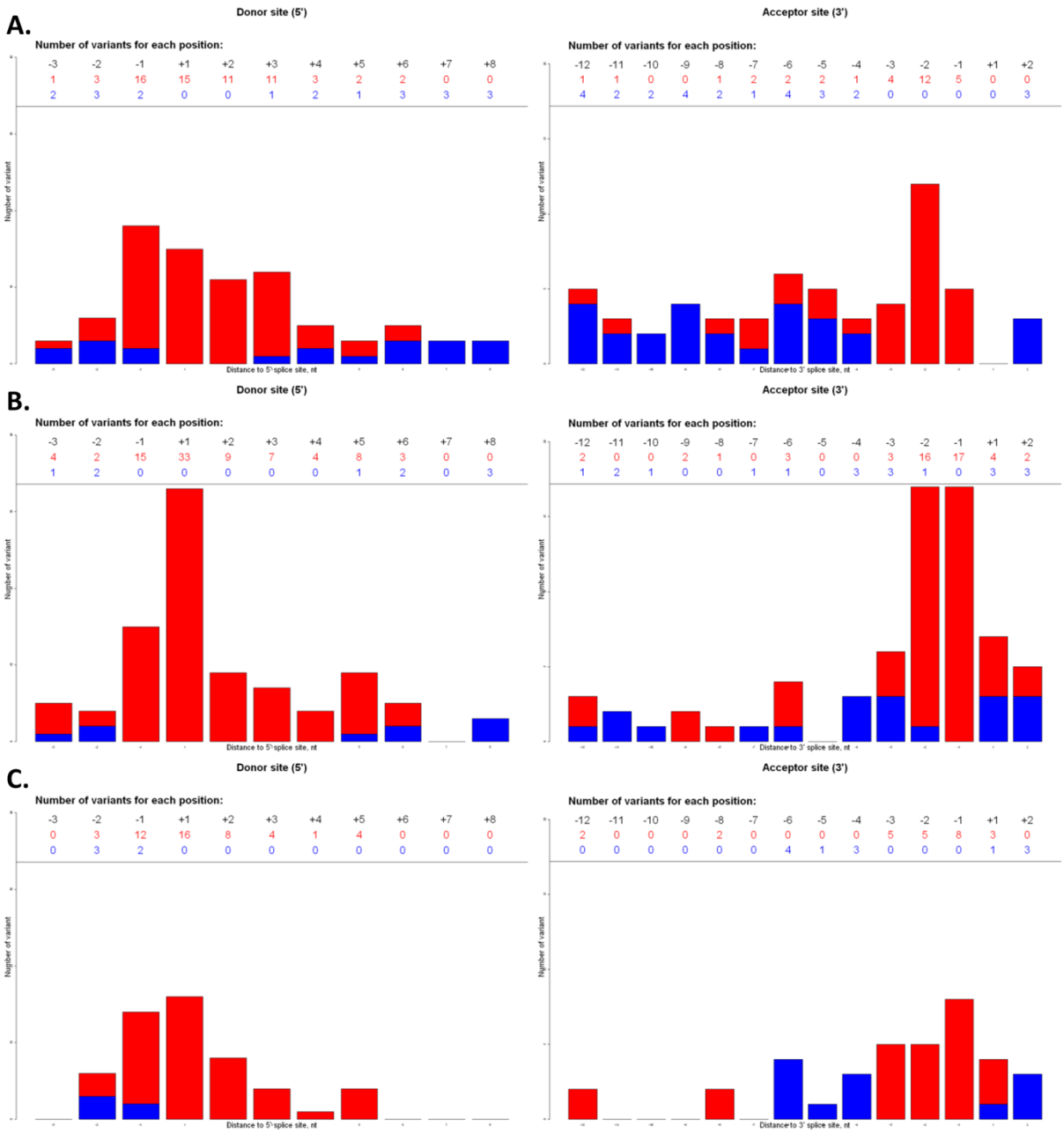


Figure 40 : Localization and impact of variants according to distance from splice site. X-axis: variant altering splicing (red bar), variant without effect (blue bar). Y-axis: total number of variants for each position. Donor and acceptor splice sites were defined as – 3 nt in exon to + 8 nt in intron and – 12 nt in intron to + 2 nt in exon, respectively. **A)** variants from trainning set. **B)** variants from BRCA1/BRCA2 validation set. One single base deletion that affects the canonical AG splice site does not induce aberrant splicing. The reason is that the deletion removes a "A" from the canonical "AG" but without disrupting the consensus as the following neighboring nucleotide is another "A" which in turn does preserve the consensus **C)** variants from other genes validation set.

d. Descriptive analyses of bioinformatics prediction score

To determine if prediction scores from different algorithms give similar information or not on our training set, we calculated Pearson coefficient correlation for each algorithm. The greatest correlations were between HSF and SSF-like (0.80) and between MES and NNS (0.87). GS score has the lowest

correlation with the other prediction scores (ranging from 0.43 to 0.48). Excluding GS score, the lowest values were observed between SSF-like and MES (0.71) and between NNS and HSF (0.60) (Table S5). The predictive capacity of each algorithm was measured by ROC curves. NNS and GS scores have the lowest Area Under the Curve (AUC) values (0.907 and 0.736 respectively). MES and SSF-like scores have the best and similar AUC value (0.968 and 0.952 respectively) (Figure 41). As a result, MES and SSF-like provide high predictive capacity with distinct information.

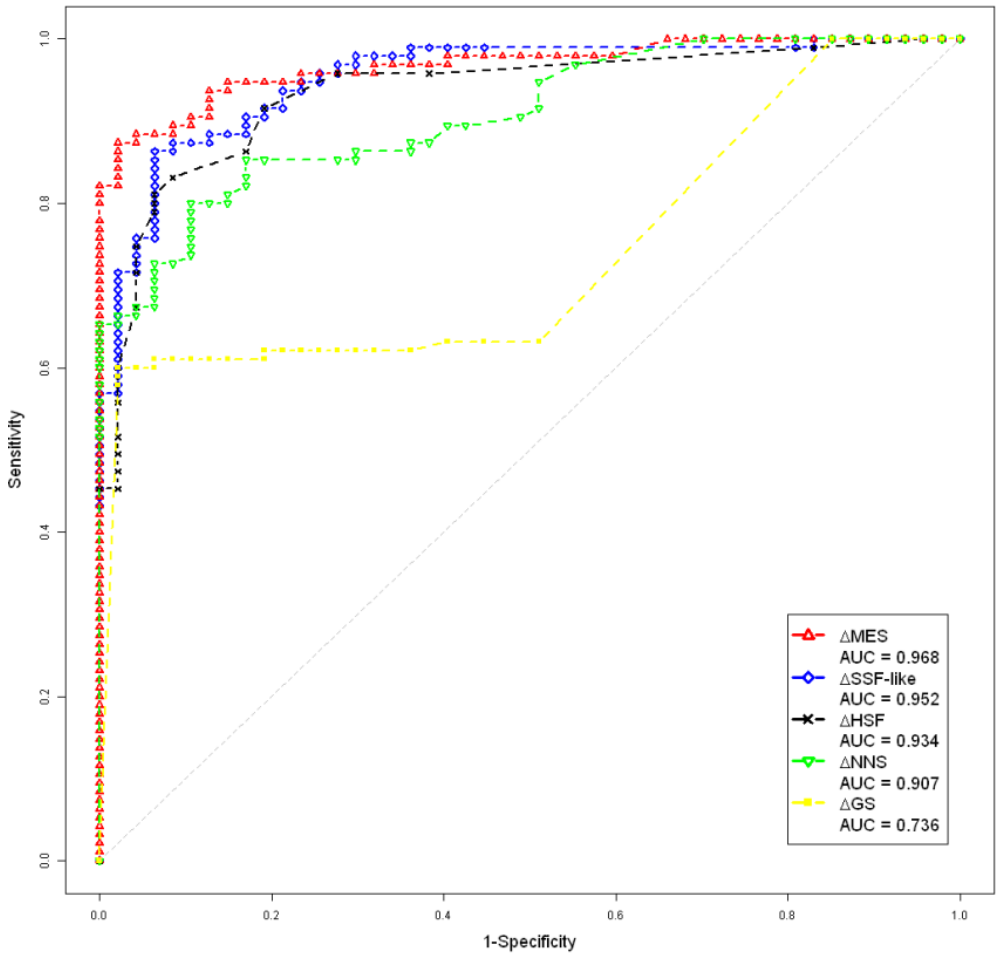


Figure 41 : ROC curves of different bioinformatics scores from the training set (n = 142). SSF: Splice Site Finder, MES: MaxEntScan, HSF: Human Splicing Finder, GS: GeneSplicer, NNS: Neural Network Splice.

e. Model definition of SPiCE

Since our last large study in 2012 [254], we collected and analyzed in the UGG network a new set of 51 variants (37 unpublished). We applied our previous guidelines to identify variant that alter splicing and obtained a sensitivity equal to 74.3% (26/35), prompting us to develop SPiCE (Table S6).

First, we performed univariate analysis for each variable (variation of prediction scores, localization in the invariant regions, donor (5') or acceptor (3') splice sites, genes). We observed that MES had a

better Akaike Information Criterion (AIC) than the other variables (63.46) (Table S7). Then, we performed multivariate analysis by adding other variables to MES. We found that only the combination of MES and SSF-like significantly improved the AIC with p-value of likelihood ratio test under 5% (Table S7). The values for intercept, MES and SSF-like parameters are shown in Table 2. These three parameters were significantly different from 0 (p-value of Wald's test < 0.05). Taken into account that MES and SSF-like don't score + 7 and +8 position of the 5'ss, SPiCE should not be used at these positions.

We determined our thresholds by using ROC curve analyses on the training set (Figure 42A). The aim of these thresholds is to prioritize *in vitro* RNA studies of variants. Two probability thresholds were thus defined: optimal sensitivity threshold (Th_{Se}) and optimal specificity threshold (Th_{Sp}), 0.115 and 0.749 respectively. As sensitivity is defined as the ratio of true positives divided by the sum of true positives and false negatives, Th_{Se} is designed to give the highest detection rate while allowing false positives. On the other hand, specificity is the ratio of true negatives divided by the sum of true negatives and false positives, meaning Th_{Sp} is designed to minimize false positives while allowing false negatives. Sensitivity and specificity with Th_{Se} are 100% (95/95) and 74.5% (35/47) respectively. Sensitivity and specificity with Th_{Sp} are 88.4% (84/95) and 95.7% (45/47) respectively. In both cases, accuracy was equal to 90.8% (data not shown). Our bootstrap analysis (Table S8 and Figure S1) confirmed stability of model parameters and thresholds. We observed that cross-validation confirmed the pertinence of combined MES and SSF-like variation scores relative to the variation scores of MES or SSF-like alone (Table S9 and Figure S2).

Table 2 : Model parameters

Parameters	Value	p-value of Wald's test
β_0	-3.59	5.48^{e}-6
β_{MES}	-8.21	4.28^{e}-3
β_{SSF}	-32.30	6.37^{e}-3

β_0: intercept, β_{MES}: parameter of MES score, β_{SSF} : parameter of SSF-like score

f. SPiCE performances on the BRCA1 and BRCA2 validation set

Following definition and training, SPiCE was validated on two independent sets of splice data. For each variant, the probability to have a splice effect was calculated and outcomes were predicted according to the previously determined thresholds (Table 3). To facilitate users' interpretation, a graphical view was developed where decision thresholds are traced and variants spotted according to their values of their SSF-like and MES score variation (Figure 43). In-between thresholds, the area is thereby defined as the "grey area" that includes only 16/160 variants. Optimal sensitivity threshold gave 99.3% sensitivity (134/135) and 68.0% (17/25) specificity. Optimal specificity threshold gave 92.6% sensitivity (125/135)

and 92.0% (23/25) specificity (Figure 42B). Accuracy values were 94.4% (151/160) and 92.5% (148/160) for Th_{Se} and $Th_{Sp,}$ respectively, *i.e*. above accuracy obtained on the training set (90.8%).

To further assess SPiCE efficiency, we compared the proportion of variants that affect splicing with their average SPiCE probability. Hence we subdivided our validation sets into groups according to their SPiCE probability. Ideally, the proportion of variants that affect splicing in any given group should be equal to the average SPiCE probability in this group. This is the case for our SPiCE model except for 4% (10/250) of variants with a probability between 0.115 and 0.432 (Figure S3).

Then we studied a possible association between prediction accuracy and distance to canonical splice site (AG/GT). As shown in Figure S4, SPiCE remains accurate throughout the consensus regions, even in the less conserved parts. However, we noted higher variability for polypyrimidine tract of 3' ss (from -5 to -12).

g. SPiCE performances on the non-BRCA validation set

For *CFTR* and *RHD* for which we tested more than 35 variants, and 7 other genes for which we tested a few variants, SPiCE classification using Th_{Se} gave a 100% sensitivity and a 82.3% specificity. Th_{Sp} gave 91.7% sensitivity and 100% specificity (Table 3). Combination of two thresholds of SPiCE protocol didn't result in misclassified variants (0 false positive and 0 false negative). These results confirmed that the SPiCE protocol is pertinent in non-*BRCA* genes.

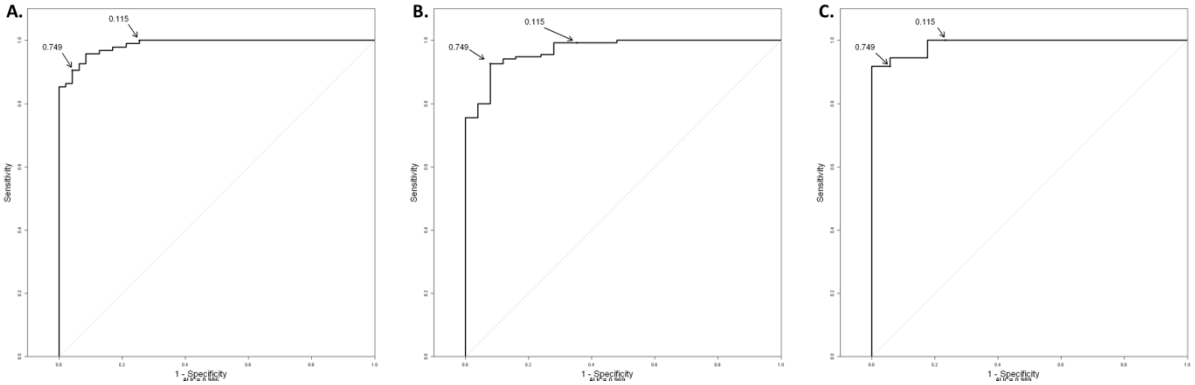


Figure 42 : ROC curve of the SPiCE logistic regression model. **A)** on training set (n = 142), AUC = 0.986. **B)** on BRCA1 and BRCA2 validation set (n = 160), AUC = 0.969. **C)** on other genes validation set (n = 90), AUC = 0.989. Arrows correspond to decision thresholds for optimal sensitivity (0.115) and optimal specificity (0.749).

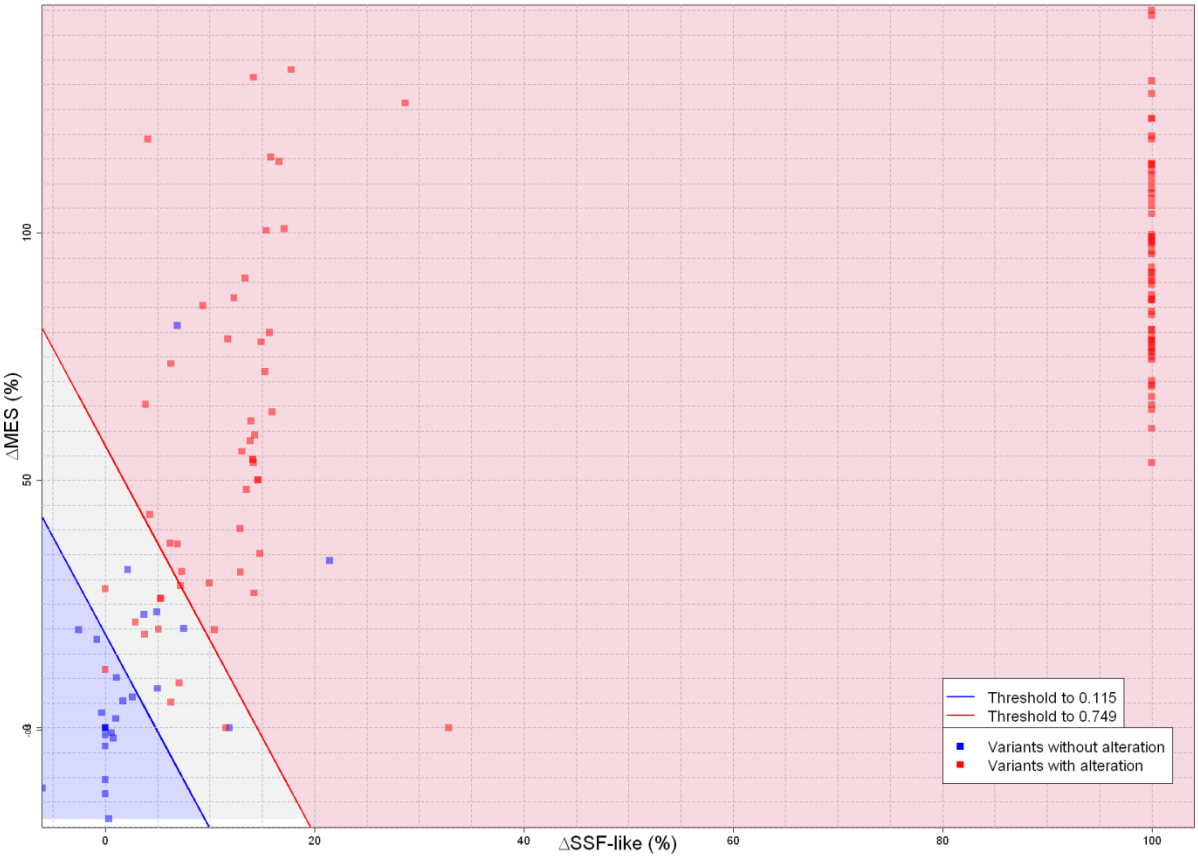


Figure 43 : SPiCE graphical ouput results on BRCA1/BRCA2 validation set (n = 160). Representation of variants according to their SSF-like and MES scores variations in percentage. Blue area represents variants with probability of splicing alteration under decision threshold of optimal sensitivity, red area corresponds to probability upper decision threshold of optimal specificity and grey area is probability between these two thresholds. Blue points are variants without splicing effect and red points are variants altering splicing.

Table 3 : SPiCE spliceogenicity prediction of variants in validation sets (n = 160 and n = 90)

	***BRCA1* and *BRCA2* validation set**		**Other genes validation set**	
	With alteration	Without alteration	With alteration	Without alteration
$P > Th_{Sp}$	125	2	67	0
$Th_{Sp} < P > Th_{Se}$	9	6	6	3
$P < Th_{Se}$	1	17	0	14

P: probability of variant to have splicing alteration, Th_{Se}: optimal sensitivity threshold, Th_{Sp}: optimal specificity threshold

h. SPiCE performances with previous published guideline

We compared the performance of our previously published guidelines to SPiCE on validation sets, n = 250 (Table 4). Using Th_{Sp}, SPiCE improves the specificity to 95.2% (40/42) against 83% with previous guidelines whereas with Th_{Se} SPiCE dramatically decreases the number of false negatives from 14 to 1 variant *i.e.* a sensitivity equals to 99.5%.

Table 4 : Contingency table on validation datasets (BRCA1/2 and other genes, (n = 250) with guidelines of Houdayer and coll. [254]

	With alteration	Without alteration
ΔMES > 15% and ΔSSF > 5%	194	4
ΔMES < 15% or ΔSSF < 5%	14	38

i. Further quantitative aspects

We questioned the capability of SPiCE to predict the quantitative nature of the splice anomalies. To this aim, 232 analyses for which the semi-quantitative effect was known were selected from the training set and the non *BRCA* validation set. These 232 analyses were for diagnostic purposes and the semi quantitative effect was taken into account for patient's reporting. As a result, and despite the well-known difficulties in splice quantification, these data were considered reliable. Semi-quantitative effect was defined using the previously published classes *i.e.* 1S (no effect on splicing), 2S (partial effect) and 3S (complete effect) [254] and plotted against SPiCE probabilities (figure S5). A trend emerged as some partial effects led to lower probabilities as compared to complete effects but we were not able to define a prediction threshold between low/high intensity effects.

5. DISCUSSION

a. General considerations

This international effort represents the largest *in silico* study of splice variants with their corresponding *in vitro/ex vivo* transcript analyses conducted to date by a consortium. These international initiatives are needed to get results of wide scale relevance *i.e*. for the whole community. It enabled us to build SPiCE, a powerful prediction tool for variants occurring at splice site consensus regions, based on combination of MES and SSF-like by logistic regression. The reason is that among the 5 algorithms tested (GS, HSF, MES, NNS, SSF-like), we found that SSF-like and MES provide the best prediction on splicing effect of variants, as previously suggested by our group and others [172]. Logistic regression analysis allows us to outperformed use of bioinformatics score variations of MES and SSF-like alone. SPiCE fulfills all the necessary criteria for model validation, *e.g*. stability of model, without bias. It has been validated on two replicative sets including 11 different genes and developed in the commonly utilized 'R' langage to ensure free and wide access.

SPiCE performs with high accuracy (95.6%) and sensitivity (99.5%) throughout the consensus sequences. The sole apparent false negative identified on *BRCA1* and *BRCA2* variants was c.5408G>C in the *BRCA1* gene that leading to exon 23 skipping. The reason for this false-negative may be due to the complexity of splicing control *i.e.* due to another mechanism, such as the disruption of distal auxiliary splicing regulatory elements. This alternative explanation could be proven by dedicated minigene assays [150], [274], [275]. Not surprisingly, there is a need for complementary prediction tools to complete our predictions and a fully comprehensive tool will eventually emerge from the combination

of SPiCE and promising splicing regulatory element predictions [274], [276], [277]. Moreover, by embedding comprehensive tools for exon definition, we would in turn be able to distinguish real exons from pseudoexons [278]. Thanks to this international network of laboratories, these novel developments are planned to address this challenge.

b. Recommendations for routine analyses

SPiCE allows the user to know the risk of missing a true splice alteration according to the probability calculated. As sensitivity is a key issue in molecular diagnosis, we would recommend using the optimal sensitivity threshold (Th_{Se}, probability above 0.115, *i.e.* including "grey area") which in our hands gave only one false negative for *BRCA1* while also a limiting number of false positives. On the other hand, depending on laboratory resources, the user can rely on the optimal specificity threshold (Th_{Sp}, probability above 0.749) which keeps false positives to a minimum as we observed only 2 false positives out of 42 variants without splice effect in our validation sets.

Previous prediction methods have been proposed for identifying variants that likely alter splicing. However, these methods were defined on small series thereby limiting their applicability [53], [259], [279]–[282]. A recent paper [283] on a large series of 272 variants in consensus regions suggested a MES threshold of relative decrease of 10% however leading to specificity of 50% (21/42) on our validation datasets. The UGG network previously published a large series of splicing variants and accompanying guidelines for *in silico* predictions [254]. Importantly SPiCE outperforms our previous results as demonstrated on the validation sets of variants from *BRCA1*, *BRCA2* and other genes (Table 3 and 4).

At this point in time, SPiCE predicts potential splicing alteration of variants at 5' and 3' ss but neither the type of the effect (exon skipping or use of alternative splice site) nor the importance of the effect (partial or total) are predicted, although the tool is able to detect a trend in the prediction severity of splicing defects (figure S5). This trend would allow to prioritize assays for those VUS predicted to have more severe effects on mRNA splicing. Importantly enough, SPiCE can be used beyond *BRCA1* and *BRCA2* and applied to other genes to guide geneticists in their daily practice. The majority of non-*BRCA* variants comes from 2 different genes (*CFTR* and *RHD*) but this should not create a bias as SPiCE runs MES and SSF which have been trained on our 20 000 protein-coding genes. Moreover we believe SPiCE versatility is demonstrated by testing these non- cancer genes *i.e*. involved in distinct pathways. This versatility is of special relevance as issues on misinterpretations and/or conflicting interpretations impact all fields of genetic diagnosis, leading to difficult situations for patients but also for health professionals. Given that 25% of clinical genetic results from commercial cancer panels had conflicting interpretation in ClinVar, the variant interpretation challenge is prone to erroneous medical decisions and eventually lawsuit as shown in Dravet syndrome [258]. Without doubt, the development of reliable *in silico* tools is a major improvement towards reliable variant classification and patient's management.

Overall, SPiCE has the potential of a widely used decision-making tool to guide geneticists towards relevant spliceogenic variants in the deluge of high throughput sequencing data.

6. DEDICATION

This work is dedicated to the memory of our colleague Olga Sinilnikova.

7. AVAILABILITY

SPiCE software is available at (https://sourceforge.net/projects/spicev2-1/)

8. SUPPLEMENTARY METHODS AND DATA

Supplementary Data are available at NAR online

9. FUNDING

This work was supported by an NHMRC Senior Research Fellowship (ID1061779) (ABS), and in part by funding from The Cancer Council Queensland (ID1086286) (MP). This work was also supported by a translational research grant from the French National Cancer Institute and the Direction Générale de l'Offre des Soins (INCa/DGOS).

10. ACKNOWLEDGMENTS

The authors wish to thank Valentin Harter for biostatistics analysis advices and Emma Tudini for gathering data from the literature.

11. CONFLICT OF INTEREST

The authors declare no conflict of interest.

II. Évaluation des outils de prédiction des points de branchement pour prédire la présence de point de branchement et leur altération par des variants : Article II

Le présent travail est en cours d'évaluation par *BMC genomics* et le pre-print est accessible à : https://dx.doi.org/10.21203/rs.2.12748/v1. Les données supplémentaires de cet article sont en **ANNEXE B**, les tableaux contenant la liste des sites d'épissage décrits par Ensembl, détectés par RNA-seq et la liste des variants avec leurs prédictions et leurs données ARN (*tables S1 to S2*) sont disponibles en ligne à https://github.com/raphaelleman/BenchmarkBPprediction.

Les points de branchement représentent de courts motifs et sont situés à une distance variable en amont des sites accepteurs d'épissage. Cette double caractéristique a rendu leur prédiction particulièrement difficile. D'autant plus que nous ne disposons d'une grande collection de points de branchement validés expérimentalement que depuis 2015. Aussi au cours de ces dernières années, plusieurs outils de prédiction ont été développés. Parmi eux nous pouvons citer HSF, SVM-BPfinder, BPP, Branchpointer, LaBranchoR et RNABPS. De plus, des variants impliqués en maladies humaines ont été identifiés comme splicéogénique en abolissant un point de branchement [51]. Par conséquent la prédiction de ces points de branchement et notamment de leur altération représente un enjeu majeur pour le diagnostic moléculaire. Pour ce travail nous avons ainsi évalué et comparé les performances de ces outils sur trois jeux de données.

L'utilisation d'un site accepteur est conditionnée par la présence d'un point de branchement en amont. Aussi un premier jeu de données contenait la liste des sites d'épissage accepteurs déduits des transcrits décrits dans la base de données Ensembl (regroupant les structures de nombreux transcrits) ainsi que la liste des sites accepteurs putatifs (tout motif AG hors ceux listés précédemment).

Le second jeu de données regroupait les sites accepteurs alternatifs identifiés par RNA-seq dans notre équipe Inserm U1245. Ces sites alternatifs ont été déterminés par rapport aux transcrits pleine longueur décrits dans RefSeq. Pour chaque site alternatif, l'expression relative a été calculée entre les transcrits supportant ces sites et les transcrits pleine longueur.

Le troisième jeu de données correspondait à une collection de variants situés dans la région d'un point de branchement avec leurs études ARN *in vitro* issue de la littérature et d'une collaboration internationale *via* le consortium ENIGMA.

Le jeu des données issu d'Ensembl nous a permis de tester la capacité des outils à détecter des points de branchement relevants. Ces derniers ont été définis comme étant un point de branchement situé en amont des sites accepteurs décrits dans Ensembl. Les données de RNA-seq ont permis d'évaluer la corrélation entre la présence d'un point de branchement prédit par l'outil et le niveau relatif d'expression du site accepteur alternatif. Enfin, la collection de variants a permis d'évaluer les performances de ces outils pour prédire l'altération d'un point de branchement par un variant.

Ainsi, l'outil Branchpointer s'est avéré particulièrement efficace pour détecter des points de branchement relevants avec une exactitude de 99.48 %. En ce qui concerne les sites accepteurs alternatifs, Branchpointer a également montré une meilleure performance que les autres outils avec une exactitude de 65.8 %. Cependant pour un certain nombre de sites accepteurs alternatifs, l'outil ne détecte pas de point de branchement. Aussi la sensibilité de l'outil décroit de 95.54 % à 32.1 % entre les données Ensembl et les données RNA-seq. Ces résultats peuvent être expliqués par le fait que Branchpointer a été entrainé uniquement sur les points de branchement qualifiés de *high-confidence*, c'est-à-dire dont le niveau d'expression des transcrits a pu permettre la confirmation de ces points de branchement par l'approche RNA-seq lasso [15].

Concernant la prédiction de l'altération des points de branchement par un variant, l'outil BPP s'est révélé être optimal pour détecter cette altération, avec une exactitude de 89.17 %. Nous avons démontré qu'un variant situé dans le motif tétramérique TRAY du point de branchement, permet une prédiction optimale d'un défaut d'épissage. Etant donné la courte séquence des motifs des points de branchement, un variant situé dans ces motifs a une probabilité élevée d'en altérer la séquence notamment au niveau du A de branchement et du T en -2. En revanche, la variation du score associé au point de branchement entre les séquences sauvages et mutées, s'avère non-optimale pour prédire un variant splicéogénique. En effet en plus du motif des points de branchement, les outils de prédiction requièrent la connaissance de l'environnement nucléotidique au-delà du motif pour pouvoir attribuer un score au point de branchement. Ainsi un variant peut modifier le score d'un point de branchement sans en altérer le motif de ce dernier.

Nous avons ainsi pu montrer d'une part que l'outil Branchpointer permet de détecter des points de branchements relevants et d'autre part que l'outil de prédiction BPP permet d'aider les généticiens à prioriser les études ARN pour les variants situés dans ces régions de points de branchement.

Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants

Raphaël Leman[†1,2,3+#*], Hélène Tubeuf[†2,4], Sabine Raad[2], Isabelle Tournier[2#], Céline Derambure[2], Raphaël Lanos[2], Pascaline Gaildrat[2+#], Gaia Castelain[2], Julie Hauchard[2], Audrey Killian[2], Stéphanie Baert-Desurmont[2], Angelina Legros[1], Nicolas Goardon[1,2], Céline Quesnelle[1], Agathe Ricou[1,2], Laurent Castera[1,2], Dominique Vaur[1,2], Gérald Le Gac[5], Chandran Ka[5], Yann Fichou[5], Françoise Bonnet-Dorion[6+], Nicolas Sevenet[6+], Marine Guillaud-Bataille[7+], Nadia Boutry-Kryza[8+], Inès Schultz[9+], Virginie Caux-Moncoutier[10+], Maria Rossing[11#], Logan C. Walker[12#], Amanda B. Spurdle[13#], Claude Houdayer[2+#], Alexandra Martins[2+#], Sophie Krieger[1,2,3+#*]

[†]These authors contributed equally to this work.
Unicancer Genetic Group (UGG) splice network[+] and ENIGMA[#]

[1]Laboratoire de Biologie Clinique et Oncologique, Centre François Baclesse, France, [2]Inserm U1245, Normandy Center for Genomic and Personalized Medicine, Rouen, UNIROUEN, Normandy University, France, [3]Université Caen-Normandie, France, [4]Interactive Biosoftware, Rouen, France, [5]Inserm UMR1078, Genetics, Functional Genomics and Biotechnology, Université de Bretagne Occidentale, Brest, France, [6]Inserm U916, Département de Pathologie, Laboratoire de Génétique Constitutionnelle, Institut Bergonié, Bordeaux, France, [7]Service de Génétique, Institut Gustave Roussy, Villejuif, France, [8]Lyon Neuroscience Research Center–CRNL, Inserm U1028, CNRS UMR 5292, University of Lyon, Lyon, France, [9]Laboratoire d'Oncogénétique, Centre Paul Strauss, Strasbourg, France, [10]Service de Génétique, Institut Curie, Paris, France, [11]Centre for Genomic Medicine, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark, [12]Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand, [13]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Herston, Queensland, Australia

*To whom correspondence should be addressed. Email: r.leman@baclesse.unicancer.fr, S.KRIEGER@baclesse.unicancer.fr
Present address:
Sophie Krieger, Laboratoire de biologie et génétique des cancers, Centre François Baclesse, Caen, France

1. ABSTRACT

Branch points (BPs) map within short motifs upstream of acceptor splice sites (3'ss) and are essential for splicing of pre-mature mRNA. Several BP-dedicated bioinformatics tools, including HSF, SVM-BPfinder, BPP, Branchpointer, LaBranchoR and RNABPS were developed during the last decade. Here, we evaluated their capability to detect the position of BPs, and also to predict the impact on splicing of variants occurring upstream of 3'ss. We used a large set of constitutive and alternative human 3'ss collected from Ensembl (n = 264,787 3'ss) and from in-house RNAseq experiments (n = 51,986 3'ss). We also gathered an unprecedented collection of functional splicing data for 120 variants (62 unpublished) occurring in BP areas of disease-causing genes. Branchpointer showed the best performance to detect the relevant BPs upstream of constitutive and alternative 3'ss (99.48 % and 65.84 % accuracies, respectively). For variants occurring in a BP area, BPP emerged as having the best performance to predict effects on mRNA splicing, with an accuracy of 89.17 %. Our investigations revealed that Branchpointer was optimal to detect BPs upstream of 3'ss, and that BPP was most relevant to predict splicing alteration due to variants in the BP area.

2. BACKGROUND

Pre-mRNA splicing by the spliceosome is essential for maturation of mRNA. Moreover, splicing plays a crucial role for protein diversity in eukaryotic cells [284]. This process, named alternative splicing, produces several mRNA molecules from a single pre-mRNA molecule and concerns approximately 95 % of human genes [30]. RNA splicing requires a mandatory set of splicing signals including: the splice donor site (5'ss), the splice acceptor site (3'ss) and the branch point (BP) site. The 5'ss defines the exon/intron junction at the 5' end of each intron with two highly conserved nucleotides, mainly GT. The 3'ss delineates the intron/exon junction at the 3' end of each intron and is characterized by a highly conserved dinucleotide (mainly AG), which is preceded by a cytosine and thymidine rich sequence called the polypyrimidine tract. The branch site is a short motif upstream of the polypyrimidine tract that includes a BP adenosine, in 92 % of human BP [285]. During the first step of the splicing reaction the 2'OH of the BP adenosine attacks the first intronic nucleotide (nt) of the upstream 5'ss to form a lariat intermediate [286]. In the second step, the 3'OH of the 5' exon attacks the downstream 3'ss thereby releasing the intronic lariat and joining the two exons together.

The 5'ss and 3'ss sequences are well characterized, mostly having been experimentally mapped, which allowed the assembly of large datasets of aligned sequences [273], [287], [288]. Therefore, several reliable *in silico* tools dedicated to splice site predictions emerged, reaching an accuracy of 95.6 % [260]. In contrast, the branch sites are short and degenerate motifs that are still poorly known and difficult to predict [285]. Indeed, only the branch A and the T located 2 nucleotides (nt) upstream, are highly conserved within a 5-mer motif of CTRAY [13]. More than 95 % of BPs are located between 18 and 44 nt upstream of 3'ss [15], hereafter named the BP area. However, some BPs can be located up to 400 nt upstream of the 3'ss [289]. The identification of relevant BPs, *i.e.* BPs used by the spliceosome, represents a major challenge given the high variability of these BPs, both at localization and motif level. Disease-causing variants have most frequently been shown to be splicing motif alterations [46] and these variants can also alter BPs [51]. An accurate prediction of BP alteration represents a challenge to molecular diagnosis.

A major limit to develop accurate BP prediction tools was the limited access to experimentally-proven BPs. The first tools Human Splicing Finder (HSF) [146] and SVM-BPfinder [153] used only 14 and 35 experimentally-proven BPs in development. In 2015, a large but not comprehensive dataset of BPs was built from lariat RNA-seq experiments [15]. This collection of BPs was extended by two further studies: the first used 1.31 trillion reads from 17,164 RNA-seq data sets [290], and the second identified BPs by the spliceosome iCLIP method [291]. Thus, several bioinformatics tools for BP prediction have recently emerged: Branch Point Prediction (BPP) [261], Branchpointer [166], LaBranchoR [167] and RNA Branch Point Selection (RNABPS) [168] (**Table 5**). Briefly, HSF uses a position weighted matrix approach with a 7-mer motif as a reference (5 nt upstream and 1 nt downstream of the branch point A)

(Figure 44). SVM-BPfinder was the first to take into account, not only the branch site motif, but also the conservation of 3'ss, as well as the AG exclusion zone algorithm (AGEZ) [289] derived from the work of Smith and collaborators [292]. BPP combines the BP and 3'ss sequences and the AGEZ algorithm by a mixture model, a popular motif inference method. Branchpointer uses machine learning algorithms trained from a set of experimentally proven BPs. LaBranchoR and RNABPS are based on a deep-learning approach. LaBranchoR re-used the dataset of Branchpointer and implemented a bidirectional long short-term memory network (LSTM) shown to be performant for modeling sequential data such as natural language. RNABPS, as LaBranchoR, used the LSTM model and also implemented a dilated convolution neural network algorithm.

Here, we present a benchmarking of these six BP-dedicated bioinformatics tools on their capacity to detect a relevant BP signal and to predict a variant-induced BP alteration. The resolution of the first issue allowed highlighting the specificity of each tool, *i.e*. the identification of BPs among background noise. For this part, we used two sets of data: a large set of 3'ss described in Ensembl database and a series of alternative 3'ss observed in RNA-seq experiments. The detection of BP alteration by a variant represents also a challenge for molecular diagnostics. To this end, we used an unprecedented collection of human variants (within the BP area) with their *in vitro* RNA studies to assess the prediction of variant effect on BP function.

Figure 44 : Illustration of position weight matrix used by HSF [146]

Table 5 : Bioinformatics tools for branch point analyses, Human Splicing Finder (HSF), SVM-BPfinder, Branch Point Prediction (BPP), Branchpointer, LaBranchoR, RNA Branch Point Selection (RNABPS), with their main features and their accessibility.

Tools	Features	Accessibility	References
HSF	• Position weighted matrix • Calculation delta score from mutation • Available as a web-application	http://www.umd.be/HSF3/	[146]
SVM-BPfinder	• Support vector machine • Scan a sequence to find a branch point • Available as a web-application + Perl script	http://regulatorygenomics.upf.edu/Software/SVM_BP/	[153]
BPP	• Mixture model • Scan a sequence to find a branch point • Available as a python script	https://github.com/zhqingit/BPP	[261]
Branchpointer	• Machine learning • Detect BP from genomic coordinates • Available as an R Bioconductor package	https://www.bioconductor.org/packages/release/bioc/html/branchpointer.html	[166]
LaBranchoR	• Deep learning • Scan a sequence to find a branch point • Available as a python script + UCSC genome browser	http://bejerano.stanford.edu/labranchor/	[167]
RNABPS	• Deep learning • Scan a sequence to find a branch point • Available as a web-application	https://home.jbnu.ac.kr/NSCL/rnabps.htm	[168]

3. RESULTS

a. Bioinformatic detection of branch points among the physiological and alternative splice acceptor sites

In this study, two sets of 3'ss data were used, natural 3'ss described in Ensembl dataset and alternative 3'ss with their expression data from RNA-seq analyses.

We first retrieved 264,787 physiological 3'ss from the Ensembl data. Adding to these 3'ss, 114,603,295 control AG were used as control data (see the "Methods" section for details). Thus, we collected 114,868,082 3'ss. ROC curve analysis was then performed for SVM-BPfinder, BPP, LaBranchoR and RNABPS on the set of physiological 3'ss, as illustrated in Figure 45A. Table 6 shows the levels of accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) derived from these ROC curve analyses. In terms of the area under the curves (AUC), the score provided by BPP exhibited the best performance (AUC = 0.818). However, Branchpointer presented the highest performances with an accuracy of 99.49 % and PPV of 30.06 %. Thus, Branchpointer was the most stringent of the bioinformatic tools for detecting putative BPs upstream of natural 3'ss. Indeed, SVM-BPfinder, BPP, LaBranchoR and RNABPS detected putative BPs for each natural 3'ss and control AG. For these 4 tools, the best accuracy to distinguish natural 3'ss from control AG was reached by BPP (75.23 %). Overall, 74,539,834 3'ss had a BP predicted by at least one tool (Supplementary Figure S3). The maximal overlap of the predicted BPs was observed between LaBranchoR and RNABPS, 8.6 % (6,405,101/74,539,834 3'ss). The percentage of 3'ss with BP predicted by the five tools was 0.15 % (111,937/74,539,834). Seventy-five percent (83,892/111,937) of these 3'ss were natural splice sites.

Among the alternative junctions of whole transcriptome analysis, 51,986 alternative 3'ss were identified (see the "Methods" section for details), to which we added the same number of control 3'ss. In all, we had 2 subsets of 51,986 (103,972) acceptor sites for whole transcriptomic data (Supplementary Table S1). From these data, Branchpointer outperformed all tested tools for detecting putative BPs (Table 7). Indeed, the AUC of the three tools, SVM-BPfinder, BPP, LaBranchoR and RNABPS, did not perform above 0.612 (RNABPS) (Figure 45B). Branchpointer showed the best accuracy of 65.8 % on the alternative splice sites. Furthermore, this tool demonstrated a similar specificity with the Ensembl and RNA-seq data, 99.6 % and 99.5 %, respectively. However, on the whole transcriptome data, the sensitivity decreased by more than 60 % (from 95.5 % to 32.1 %) (Table 6 and Table 7). The alternative 3'ss and non-natural 3'ss had BPs predicted by at least one of the tools in 91.2% (94,806/103,972). The maximal overlap was observed between the four tools SVM-BPfinder, BPP, LaBranchoR and RNABPS (7,227 3'ss). More than 95 % of 3'ss with a BP predicted by Branchpointer were alternative splice sites (Supplementary Figure S4).

We compared the expression of alternative sites, from RNA-seq data, with and without the presence of a putative BP predicted by the bioinformatic tools (see the "Methods" section for details). This analysis

revealed that 3'ss with a predicted BP were significantly more expressed than 3'ss without a predicted BP, regardless of the bioinformatics tool (Figure 46). The greater difference of expression was observed for Branchpointer. The average expression was 34.00 % and 1.35 %, for alternative 3'ss with Branchpointer-predicted BP or not, respectively. In the subgroup of 3'ss with a predicted BP, the Branchpointer score was not correlated with the expression of these sites (R^2 =0.00001, p-value = 0.24). The other bioinformatics tools presented a weak correlation between their score and the expression (Supplementary Figure S5). Among SVM-BPfinder, BPP, LaBranchoR and RNABPS, the best correlation was obtained with RNABPS (determinant coefficient (R^2) = 0.0062, p-value = $4.14x10^{-70}$).

Table 6 : Performance of tools derived from contingency table with Ensembl dataset (n = 114,868,082).

	SVM-BPfinder	BPP	Branchpointer	LaBranchoR	RNABPS
Cutoff	0.706	5.384	-	0.653	0.653
TP	166,135	198,708	252,967	171,511	193,430
FP	36,526,998	28,315,554	583,920	40,370,908	30,878,750
TN	72,145,972	86,003,592	114,019,375	74,232,290	83,724,448
FN	84,113	65,422	11,820	93,276	71,357
Missing data	5,944,864	284,806	0	97	97
AUC	0.728	0.819	-	0.711	0.811
Accuracy	66.39 %	75.23 %	99.48 %	64.77 %	73.06 %
Sensitivity	66.39 %	75.23 %	95.54 %	64.77 %	73.05 %
Specificity	66.39 %	75.23 %	99.49 %	64.77 %	73.06 %
PPV	0.45 %	0.70 %	30.23 %	0.42 %	0.62 %
NPV	99.88 %	99.92 %	99.99 %	99.87 %	99.91 %

TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative), AUC (Area Under the Curve), PPV (Positive Predictive Value), NPV (Negative predictive value).

Table 7 : Performance of the bioinformatics tools on the alternative acceptor splice sites (n = 103,972).

	SVM-BPfinder	BPP	Branchpointer	LaBranchoR	RNABPS
Cutoff	0.76997	5.55569	-	0.66239	0.6962
TP	28,990	29,953	16,671	29,346	29,320
FP	22,608	22,033	206	22,640	21,894
TN	29,132	29,953	51,780	29,346	30,092
FN	22,499	22,033	35,315	22,640	21,274
Missing data	743	0	0	0	1,482
AUC	0.595	0.591	-	0.592	0.612
Accuracy	56.3 %	57.6 %	65.8 %	56.4 %	57.9 %
Sensitivity	56.3%	57.6 %	32.1 %	56.4 %	57.9 %
Specificity	56.3%	57.6 %	99.6 %	56.4 %	57.9 %

TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative), AUC (Area Under the Curve).

b. Bioinformatic prediction of splicing effect for variants in the branch point area

The last set of data was a collection of experimentally characterized potentially spliceogenic variants mapping within BP areas (see the "Methods" section for details), n = 120 variants among 86 introns in 36 different genes (Supplementary Table S2). Part of this collection was obtained from unpublished data

(n = 62 variants). From the 120 variants, 38 (31.7 %) were found to induce splicing alteration, and were therefore considered as spliceogenic, whereas 82 (68.3 %) did not show splicing alterations under our experimental conditions. Figure 47 indicates the repartition of the 120 variants within the corresponding BP areas and their impact on RNA splicing. The 38 spliceogenic variants were identified in 30 different introns; 22 variants induced exon skipping, 10 variants caused full intron retention and six remaining variants activated the use of another cryptic 3'ss located up to 147 nt upstream of the 3'ss and 38 nt downstream of the initial acceptor site (Supplementary Table S2).

After the prediction of BPs for each intron affected by the variants, we analyzed the distribution of each variant according to the position of the predicted BP (Supplementary Figure S6). First, we assayed the different size motifs to classify variants (see the "Methods" section for details). The best common motif was the 4-mer starting 2 nt upstream of the A and 1 nt downstream (Supplementary Figure S7), that corresponds to the motif TRAY. For this size motif, BPP presented the best accuracy with 89.17 % and LaBranchoR had the lower performance with an accuracy of 78.33 % (Table 8). Branchpointer did not predict a BP for the intron 24 of *BRCA2* gene causing a missed data point, corresponding to *BRCA2* c.9257-18C>A variant.

As shown in Supplementary Figure S6, variants affecting splicing were mostly located at putative branch point positions 0 (the predicted branch point A) and -2 (the T nucleotide 2 nt upstream of the branch point A itself). BPP pinpointed the highest number of spliceogenic variants in these positions. More precisely, splicing anomalies were detected for all of the ten variants occurring at position -2, and for 15 out of 18 variants predicted to be located at the branch point A. The three remaining variants predicted by BPP to alter the branch point A position (*BRCA1* c.4186-41A>C, *MLH1* c.1668-19A>G and *RAD51C* c.838-25A>G), and not experimentally validated, were also predicted to alter a BP adenosine by SVM-BPfinder while Branchpointer and LaBranchoR placed these variants outside BP motifs.

Next, we assessed the discriminating capability of each tool, including HSF, by calculating delta scores, to identify splicing defects from BP variants (Figure 45C). In terms of delta score, SVM-BPfinder outperformed the other tools with an AUC of 0.782. From this ROC analysis, we identified an optimal decision threshold (see the "Methods" section for details) of -0.136, *i.e.* the variants were predicted as spliceogenic if the variant score was less than 13.6 % of the wild-type score. The performances achieved with this threshold are reported in Table 9. SVM-BPfinder reached the maximal accuracy of 81.67 %.

The achievement of cross-validation, from the logistic regression model, highlighted the performance of combination of the BPP and Branchpointer tools (see the "Methods" section for details). This model was to infer variants as spliceogenic if they occurred within a TRAY 4-mer BP motif predicted by both BPP and Branchpointer. Although this combination was mostly found in the 1,000 simulation, this model appeared in only 26 % of these simulations (see Supplementary Figure S8). The likelihood ratio test between this model and a model with only the BPP tool was not systematically significant, with

60.1 % of simulations having p-value above 1 %. This approach also showed that for a variant in intron with different and non-overlapping predicted BP sites by BPP and Branchpointer, the model could not provide prediction of potential spliceogenicity. We continued the cross-validation without the positions of predicted BP for all tools except BPP. However, the delta scores of other tools did not improve the model, as the majority of simulations converging to BPP-alone model (Supplementary Figure S9). Thus, the analysis revealed that the position of the BPs predicted by BPP alone was the optimal model.

Table 8 : Classification of variants according their position in the predicted branch point (n = 120) (Motif 4-mer: TRAY).

	SVM-BPfinder	BPP	Branchpointer	LaBranchoR	RNABPS
TP	24	32	32	27	30
FP	6	7	12	15	12
TN	76	75	69	67	70
FN	14	6	6	11	8
Accuracy	83.33 %	89.17 %	84.87 %	78.33 %	83.33 %
Sensitivity	63.16 %	84.21 %	84.21 %	71.05 %	78.95 %
Specificity	92.68 %	91.46 %	85.19 %	81.71 %	85.37 %

TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

Table 9 : Contingency table of variant according to the variation score, n = 120 variants.

	HSF	SVM-BPfinder	BPP	Branchpointer	LaBranchoR	RNABPS
Cutoff	-0.0378	-0.136	-0.0006	-0.0003	-0.0194	-0.0304
TP	27	29	22	10	25	27
FP	18	13	31	13	25	20
TN	62	69	51	59	57	62
FN	12	9	16	16	13	11
AUC	0.750	0.782	0.638	0.645	0.710	0.763
Accuracy	75.4 %	81.67 %	60.8 %	70.4 %	68.3 %	74.2 %
Sensitivity	71.1 %	76.32 %	57.9 %	38.5 %	65.8 %	71.1 %
Specificity	77.5 %	84.15 %	62.2 %	81.9 %	69.5 %	75.6%

TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative), AUC (Area Under the Curve).

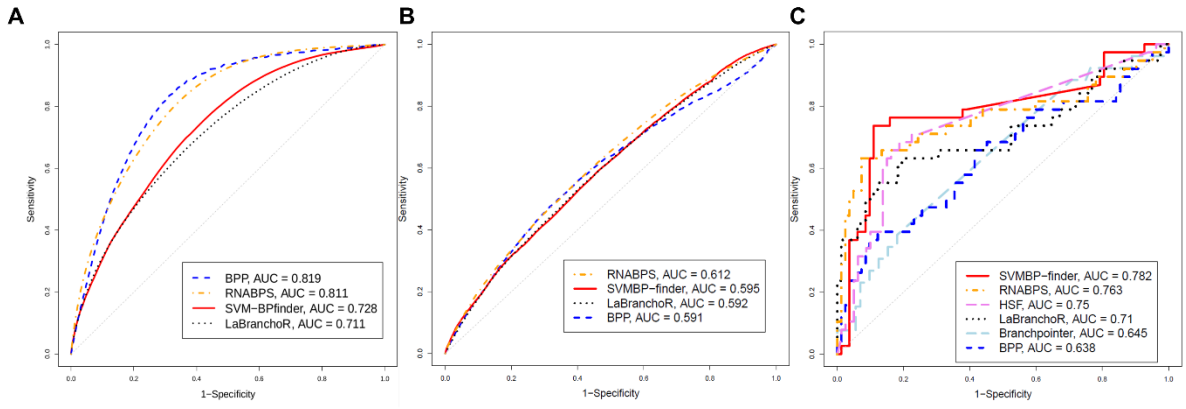


Figure 45 : ROC curves of the bioinformatics scores. For each possible score threshold, sensitivity and specificity were plotted. **A.** The detection of branch points from the set of physiological acceptor

splices sites (n = 114,868,082) of BPP, SVM-BPfinder, LaBranchoR and RNABPS scores. **B.** The detection of branch points from the alternative 3'ss by the SVM-BPfinder, BPP and LaBranchoR (n = 103,972). **C.** The delta scores of HSF, SVM-BPfinder, BPP, Branchpointer, LaBranchoR and RNABPS to class variants (n = 120).

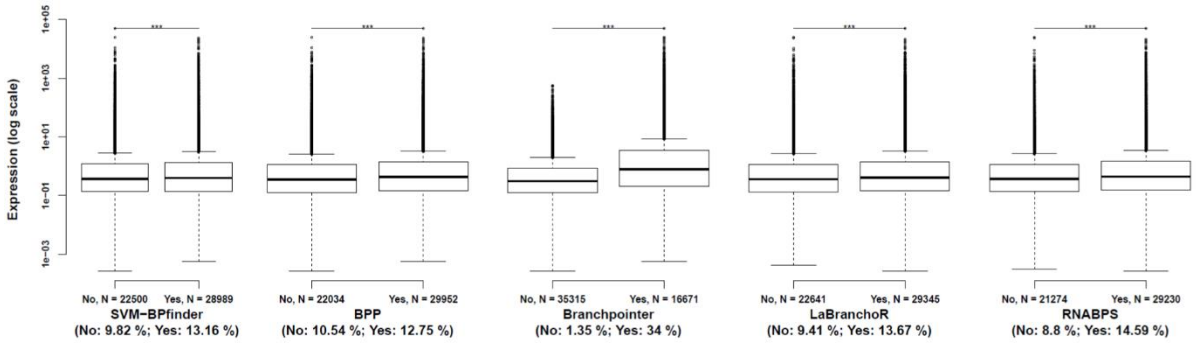


Figure 46 : Expression of 3'ss according the presence or not of predicted branch point by the bioinformatics tools, from RNA-seq data (n = 51,986 3'ss). ***: p-value (Student test) <2e-16. In brackets, the average expression between the two groups.

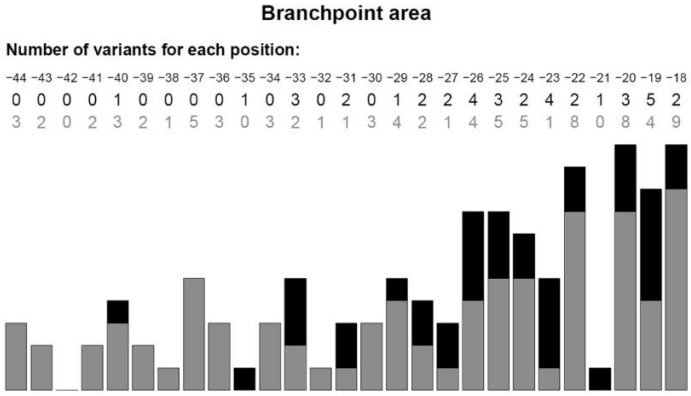


Figure 47 : Distribution of intronic variants in the branch point area (-18 to -44) experimentally tested for their impact on RNA splicing (n = 120). Positions are relative to the nearest reference 3'ss. In black variants that altered RNA splicing. In grey, variant without effect.

4. DISCUSSION

In this study we benchmarked 6 different tools for their ability to detect either a physiological BP, or a variant-induced BP alteration. From physiological data, Branchpointer showed the best performance with an accuracy of 99.48 %. This highlighted the interest of the machine learning approach compared to support vector machine and mixture models used in the development of SVM-BPfinder and BPP, respectively. One would have expected that RNABPS and LaBranchoR, using a deep learning approach, should have performance equal or above to Branchpointer. However, on the physiological set, these tools reached an accuracy of 64.77 % (LaBranchoR) and 73.06 % (RNABPS) (Table 6). This unexpected results might be explained by the fact that Branchpointer takes into account the structure of transcript unlike LaBranchoR and RNABPS. The relative expression of junctions was significantly correlated to the bioinformatic scores. However, these correlations remain weak, with a maximum coefficient of

determination (R^2) of 0.0062 for RNABPS. Added to this, even if Branchpointer had shown the best performance, the sensitivity of Branchpointer decreased by almost 60 % (95.54 % to 32.1 %) between the physiological and alternative splicing datasets. Alternative 3'ss, without Branchpointer prediction, were expressed at relative low levels. Branchpointer was trained on the high-confident BPs and the low confidence BPs were considered as negative [166]. This issue highlighted the limit of detection of Branchpointer, for the weakly used 3'ss or the less conserved BPs. The performance of Branchpointer confirms the importance of the BP in 3'ss definition, but does not explain the expression level of these 3'ss. This last point highlights the complexity of splicing that does not only depend on the 5'ss, 3'ss and the BP. To illustrate this complexity, a recent study was published [293] demonstrating the MMSplice tool which gathers several features from intronic and exonic pre-mRNA sequences. This tool was assayed on the Vex-seq data [100] which consists of 2,059 human genetic variants in and around 110 exons. For each variant the authors displayed the percentage of exon inclusion by minigene splicing assays. The correlation between this percentage and the MMSplice score reached an R^2 of 0.48 ($=0.69^2$). Despite accounting for both set of splicing motifs and the BP motifs, more than 50 % of expression variability of exon inclusion remained unexplained by the predictions.

For the variants occurring in the BP area, we gathered a large collection of 120 human variants (62 unpublished), with their corresponding *in vitro* RNA data. From our analysis, the best prediction strategy was to consider the variant as impacting the splicing if it is located in the BP motif. With this strategy the best score was obtained by BPP with an accuracy equal to 89.17%. We observed that only 31.7% (38/120) of variants altered the splicing in the BP area while 82.05 % (32/39) alter splicing in the BPP-predicted BP motif with a sensitivity of 84.21 %. These results demonstrate the interest of BPP for prioritization of variants occurring in this region for molecular diagnostic laboratories. From our dataset, we first determined, that the 4-mer **T**R**A**Y in the BP motif was the most impacted by variants. A variant occurring in this motif has a high probability to alter splicing. In our work, this probability was 82 % with BPP tool while the proportion of variants affecting the splicing outside this motif was 7.4 %. This bioinformatic tool takes into account several features in and around the 4-mer motif. Variants outside the BP motif can modify the score of BPs, although having a weak risk splicing alteration. Thus variants can wrongly affect the score. Indeed, 37 (45.7 %) variants occurring outside this 4-mer motif decreased the BPP score whereas only 4 (10.8 %) of these variants impacted splicing. Therefore, we excluded the delta score used to predict the BP alteration by a variant. The alignment of variants on the BPP-predicted BP revealed that the most spliceogenic variants were localized at the nucleotide position 0 (A) and -2 (T) of the BPs. The highly conserved di-nucleotides at the position 0 and -2 [294] were critical to the BP recognition. Thus, the BPs predicted by the BPP tool seemed to be relevant. The first study of a large collection of BPs identified the presence of redundant BPs [15]. We also observed that variants altering high BP scores, as predicted by BPP induced splicing alterations in the vast majority of cases (82 %). Among the introns (n = 86) studied in this work, the potential redundancy of BPs was not sufficient to

allow natural splicing to be completely restored. In our analyses, we did not focus on the quantitative effect of splicing, due to the diversity of RNA *in vitro* studies. Among the data generated in this study, eight of the variants that impacted splicing were assessed using minigene assays. In these condition, these variants produced both the natural and aberrant transcripts, *i.e.* they had a partial effect (data not shown). The presence of redundant BPs could explain this partial effect. However, this was beyond the scope of the present benchmarking study and will need to be explored in future studies. We observed that sequence alteration of BPs induced not only exon skipping but also intron retention and the use of new distant 3'ss. Thus, these predictions will permit the prioritization of RNA *in vitro* studies rather than determine the exact effect on splicing. The combination of BPP and Branchpointer, slightly improved prediction of BP position. Moreover, for introns with non-overlapping BPP and Branchpointer-predicted BP positions, the model will not conclude the spliceogenicity of a variant. From a practical view point the combination of scores makes the predictions of BPs less accessible.

The accessibility of the tools represents a technical limit to the analysis of BP. Indeed, HSF, SVM-BPfinder and RNABPS have a Graphical User Interface web page for non-bioinformatician users. However, LaBranchoR and BPP score calculation was only accessible by a python script. LaBranchoR also offers a list of potential BPs predicted by the tool and visualization via the UCSC Genome Browser [108]. Branchpointer is only accessible by an R package and needs the installation of several other libraries. Due to machine learning calculation, this tool also has the longest run-time. The score calculation for the physiological data set (n = 114,868,082) with a Linux machine AMD® Ryzen 7 pro 1700 eight-core processor, 8 Gb of RAM with multiprocessing way (6 at the same time) took more than two weeks, instead of a couple of days for SVM-BPfinder. Added to this, HSF tool did not allow an analysis of batches of the BP and so makes the analysis difficult of variant obtained from next-generation sequencing.

5. CONCLUSION

Our study spotlighted the requirement to distinguish two issues, the capacity to detect a real BP and the capacity to predict the splicing alteration at BP level. Branchpointer exhibited the best performance to detect a real BP from our physiological and alternative 3'ss datasets. For research purposes, Branchpointer facilitates the study of alternative transcripts by predicting the most likely used alternatively spliced 3'ss. However, the BPP-predicted BPs were more efficient to predict the impact of variants on BP usage. Furthermore, BPP was able to predict 4-mer BP motifs, with an accuracy of 89.17 %. Using a large collection of human variants (n = 120) with associated RNA *in vitro* splicing data, we confirm the advantage of studying the BP area ([-44 -18] intronic positions) for application to molecular diagnostics. As the next generation sequencing era increases the number of variants detected across exonic and intronic regions, we show how these BP prediction tools can assist the diagnostician by prioritizing variants for *in vitro* RNA studies.

6. METHODS

a. Sets of data

The Ensembl dataset contains the coordinates of a large collection of transcripts [295], with more than 200,000 human transcripts; both physiological and major alternative transcripts (download June 28th 2018). We extracted the position of exons for each described transcript then we deduced the coordinates of splice sites. As control data, we took all AG sequences found in each transcript sequence, named hereafter control AG. For each 3'ss, the genomic coordinates were annotated according to the hg19 genome assembly.

We defined as alternative splice sites all 3'ss identified from RNA-seq data that were not described in the transcripts from the RefSeq dataset [6] The alternative 3'ss were obtained from our in-house RNA-seq analyses. The read count mapped on these last RefSeq 3'ss served as a reference to calculate the relative expression of the alternative 3'ss. Whole transcriptome RNA-Seq experiment was performed on 72 RNA samples corresponding to lymphoblastoid cell lines (LCLs) from four control individuals and eight patients with pathogenic variants in *TP53* or in the *BRCA1/2* genes, treated and untreated with bleomycin or doxorubicin, and performed in triplicate. Ribosomal RNA was depleted using the NEBNext® rRNA Depletion Kit (Human/Mouse/Rat) (NEB, Ipswich, MA, USA) and libraries were produced using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB). 2x75b paired-end sequencing was performed on an Illumina NextSeq500 yielding an average of 50 million paired reads per sample. Reads were aligned on the Ensembl reference genome GRCh37 release 75 (ftp://ftp.ensembl.org/pub/) using STAR v2.5.3a tool (Spliced Transcripts Alignment to Reference) [114] and counting was performed using FeatureCounts tool v1.5.2 [121]. To avoid the impact of cell culture condition and the effect of variants on the expression of alternative 3'ss, we selected alternative splice sites observed in more than six samples. The expression of alternative splice sites was calculated as follows [131]:

$$\%_{expression} = \frac{read\ count_{alternative\ site}}{read\ count_{physiological\ site}} \times 100$$

The read count corresponded to the number of reads mapping on exon junctions and the physiological site was defined as the nearest splice site, described in RefSeq, and same type of alternative splice site.

As control data for the set RNA-seq data, we took any non-natural 3'ss that had a MaxEntScan [148] score higher than 0 but was not identified as an alternative splice site. From these control data we randomly selected 3'ss so that the number of sites in the control dataset was equivalent to the number of alternative 3'ss.

The last set of data was a collection of potential spliceogenic variants, characterized by experimental RNA studies, occurring in the BP area (from -18 to -44 relative to the 3'ss) of 36 genes. Briefly, this

dataset included RT-PCR data obtained from (i) minigene-based splicing assays (by Inserm U1078 and by Inserm U1245 teams), (ii) RNA extracted from lymphoblastoid cell lines treated/untreated with puromycin, (iii) RNA extracted from blood collected into PAXgene tubes (Qiagen), (iv) RNA extracted from stimulated T lymphocytes provided by the French Splice Network of the Unicancer Genetic Group. Controls (samples a without a variant in the BP area) were systematically included in these experiments [254]. During the collection, we excluded any variant that altered splicing by creation or reinforcement of a cryptic or *de novo* consensus splice site. Owing to the fact that the data were heterogeneous in term of analyses and submitters, we did not take into account the quantitative information of splicing alteration. Thus, we pooled together variants having a partial or total effect on splicing.

b. Assessment of bioinformatics tools

Six BP-dedicated *in silico* tools were tested: HSF v3.1, SVM-BPfinder, BPP, Branchpointer v3.8, LaBranchoR and RNABPS (Table 1). On the other hand, we were confronted with an inaccessible tool, the BPS predictor [296], at the time of this work, so it was excluded from this study. HSF v3.1 did not allow browsing of a wild-type sequence to detect potential BPs, and gave only the score change of a variant. For the other tools, we used the standalone versions that were: python scripts for SVM-BPfinder, BPP and LaBranchoR and R package for Branchpointer. For the tools SVM-BPfinder, BPP and Branchpointer, we narrowed the browsed sequence by these 3 tools to include 1 to 200 nt upstream of the 3'ss. LaBranchoR and RNABPS need a 70 nt long sequence, and the browsed sequence was the 70 last nt of the intron.

Receiver operating characteristic (ROC) analysis was performed for the tools generating continuous scores (SVM-BPfinder, BPP, LaBranchoR, RNABPS). From each ROC curve, we determined an optimal decision threshold defined as the threshold with the minimal difference between the sensitivity and specificity. Branchpointer displayed only BPs with high confidence level, so we processed directly to a contingency table between the true 3'ss and the control AG with predicted BPs (Supplementary Figure S1).

From the RNAseq data, the relative expression of alternative 3'ss was studied according the BP-predictions of bioinformatic tools. We compared the expression between the two groups: 3'ss with predicted-BP and 3'ss without predicted-BP. The Student test was used under the hypothesis that the relative expression follows a log-normal distribution.

To study the effect of nucleotide variants on RNA splicing, we considered two questions, i) Is the variant located in a putative BP? and ii) Does the variant decrease the score of the putative BP? Given that the first question concerns a binary variable, we used contingency tables to compare the performance of the different tools. We started by using five out of the six tools (exclusion of HSF) to define a list of predicted BPs in the browsed sequences from each intron that are affected by the variants in our dataset. Next, we took only one BP with the highest score per intron, for each tool. To determine whether the

variant was located in the motif of the predicted BP, we assayed different motif sizes from 1 nt (corresponding to the branch point A) up to 7-mer around the A, *i.e.* the 3 nt on either side of A. The 7-mer motifs corresponded to the length of position weight matrices used by the majority of the tools (Figure 44). We established the optimal motif size as having the best compromise of sensitivity and specificity across all tools. The second question involved the calculation of a delta score defined as follows:

$$Delta_{score} = \frac{Score_{variant\ site} - Score_{wildtype\ site}}{Score_{wildtype\ site}}$$

This delta score did not necessarily imply that wild-type and variant scores were from the same BP site. Different examples are illustrated in Supplementary Figure S2. On this delta score we performed ROC curve analyses and then defined an optimal decision threshold to classify the variants.

c. Evaluation of the score combination

To determine the optimal score combination, we used a logistic regression. This model provided a probability that the variant alters RNA splicing depending on the information given by the bioinformatic scores. We performed a cross-validation, with two thirds of the data being used as training set and the remaining data as validation set. The data was allocated at random and this step was repeated 1000 times. On the training set, we executed a step-by-step variable selection (stepwise). On the validation set, the performances of the probability given by the model were evaluated by a ROC curve analysis.

7. ADDITIONAL FILES

Additional file 1: Figures S1-9

Additional file 2: Table S1. Collection of alternative acceptor splice site (3'ss) and controls AG (n = 103,972), from RNA-seq data

Additional file 3: Table S2. Collection of variants used to compare the branch point predictions (n = 120)

8. DECLARATION

a. Ethics approval and consent to participate

All subjects gave informed consent for genetic analysis and the consents were approved by the French Biomedicine Agency (https://www.agence-biomedecine.fr/).

b. Consent for publication

Not applicable

c. Availability of data and material

The scripts used for the evaluation of algorithms' performance are available at https://github.com/raphaelleman/BenchmarkBPprediction. The set of natural 3' ss and control AG were constructed with the dataset download from Ensembl [295] (https://www.ensembl.org/index.html). The alternative 3'ss, from RNAseq data, and controls were shown in supplemental (Supplemental file 2: Table S1). All variants reported in this study were in supplemental information (Supplemental file 3: Table S2).

d. Competing Interests

All authors except H.T. declare that they have no competing interests. HT was employed by Interactive Biosoftware for the time period October 2015-September 2018 in the context of a public-private PhD project (CIFRE fellowship #2015/0335) partnership between INSERM and Interactive Biosoftware.

e. Funding

We are grateful to the French *Fondation de France* (200412859), the *Institut National du Cancer/Direction Générale de l'Offre de Soins* (INCA/DGOS, AAP/CFB/CI), the *Cancéropôle Nord-Ouest* (CNO), the *Groupement des Entreprises Françaises dans la Lutte contre le Cancer* (Gefluc, # R18064EE), and the OpenHealth Institute for supporting this work. RNA-Seq experiments were funded by the *Canceropole Nord Ouest* (CNO). R.L. was co-supported to the *Fédération Hospitalo-Universitaire* (FHU), HT was funded by a CIFRE PhD fellowship (#2015/0335) from the French *Association Nationale de la Recherche et de la Technologie* (ANRT) in the context of a public-private partnership between INSERM and Interactive Biosoftware, and S.R. was co-supported by the *Ligue contre le Cancer*, *the European Union* and the *Région Normandie* (European Regional Development Fund, ERDF). LCW is supported by a Rutherford Discovery Fellowship (Royal Society of New Zealand). ABS is supported by an NHMRC Senior Research Fellowship (ID1061779).

f. Authors' contributions

R.L. initiated this work, performed bioinformatics and biostatistics analyses and wrote this paper. H.T., P.G, G.C., A.K. and J.A. provided minigene splice-assay data. S.R., R.L. C.D. and I.T. performed the RNA-seq analyses. S.B.D., A.L., N.G., C.Q., A.R., L.C., S.K., D.V., G.L.G., C.K., Y.F., F.B.D., N.S., M.G.B., N.B.K., I.S., V.C.M and M.R., provided RNA splicing data and their interpretation. L.C.W and A.B.S checked and confirmed result. C.H., A.M. and S.K. directed this study. All authors read and approved the final manuscript.

g. Acknowledgements

We wish to thank Alexandre Atkinson and Thibaut Lavole for their bioinformatics help in particular to perform the multiprocessing analysis. We also thank Tayara Hilal for his help to get the RNABPS score.

III. SPiP : un nouvel outil pour adresser à la diversité des altérations de l'épissage

Au sein de cette dernière partie du travail de thèse, nous nous proposons de détailler l'outil SPiP (*Splicing Prediction Pipeline*), que nous finalisons de développer afin de permettre la détection de variants splicéogéniques quel que soit leurs positions dans le gène et le motif affecté. De plus pour améliorer le poids des prédictions, nous avons estimé la probabilité qu'un variant altère l'épissage en fonction de sa position et des prédictions associées. L'article décrivant SPiP est en cours de rédaction et la version actuelle est accessible en **ANNEXE C**.

Chaque variant nucléotidique peut modifier l'épissage indépendamment de sa position dans le gène. Ces modifications d'épissage peuvent être regroupées en quatre catégories : le saut d'exon, l'utilisation de nouveau site d'épissage, la création de pseudo-exon et la rétention d'intron. Pour avoir un jeu de données représentatif de ces diverses altérations, nous avons collecté 2 784 variants avec études ARN *in vitro*. Cette importante collection de variants a été obtenue grâce à un recueil de la littérature et au travail du consortium ENIGMA et du réseau épissage du GGC ainsi que des collègues généticiens moléculaire de l'ANPGM. Parmi ces variants, 47 % (1 294/2 784) impactent l'épissage. La répartition des variants en fonction de leur effet est illustrée en Figure 48.

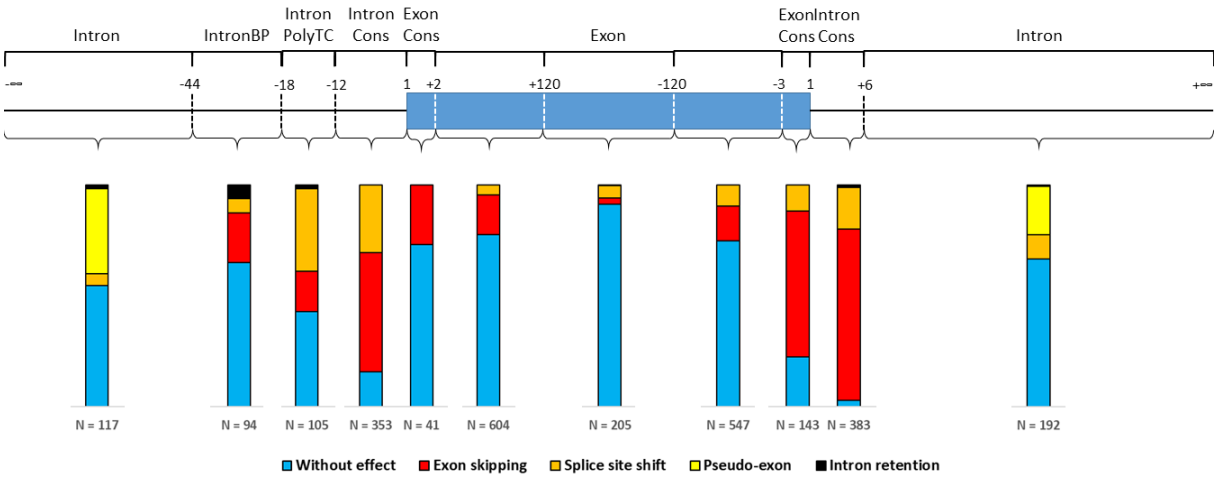


Figure 48 : Répartition des variants dans les différentes régions du transcrit, N = 2 784 variants. En bleu les variants sans effet sur l'épissage, en rouge les sauts d'exon, en orange l'utilisation d'un nouveau site d'épissage, en jaune les pseudo-exons et en noir les rétention complètes d'intron. BP : Branch point area, PolyTC : polypyrimidine tract, Cons : Consensus splice site.

SPiP est un algorithme décisionnel intégrant les outils optimaux pour chaque motif d'épissage. La définition de ces outils optimaux est basée sur nos précédents travaux ainsi que sur les données de la littérature. Ainsi SPiCE a été choisi pour l'altération des sites consensus [260]. L'outil BPP a été utilisé pour les points de branchement [262]. L'outil ΔtESRseq a été sélectionné pour prédire l'altération des

ESRs [276], [297]. Pour l'outil optimal de prédiction du tract polypyrimidique des sites accepteurs, notre choix a été guidé par le travail du groupe épissage du GGC [254].

Le tract polypyrimidique, considéré ici, est situé entre le site consensus accepteur défini par SPiCE et la région des points de branchement, soit une région entre la 13ème nt et la 17ème nt de l'intron. Le tract polypyrimidique étant étroitement associé à la séquence consensus du site accepteur, il n'existe pas d'outil dédié à ce motif. Aussi, il a été utilisé les outils dédiés aux sites consensus d'épissage, prenant en compte ce tract au moins jusqu'au 17ème nt dans l'intron. Parmi les outils répondant à ces critères, MES se révèle être un bon candidat. Car il prend en compte la séquence du site accepteur jusqu'à 20 nt dans l'intron [148] et il est l'un des outils optimaux pour prédire l'altération de cette séquence [172], [254], [260]. De plus nous avons reconfirmé *a posteriori* les performances de MES. Pour cela nous avons comparé les outils MES et SPANR sur les 65 variants situés entre la 13ème nt et la 17ème nt de l'intron, parmi les 2 784 variants de notre étude. Il en a résulté que le score MES atteint une AUC_{ROC} de 0.909 et le score SPANR une AUC_{ROC} de 0.639. Ainsi l'outil MES a été sélectionné pour le tract polypyrimidique.

Etant donné l'absence de recommandations dédiées à la prédiction de l'utilisation de nouveau site d'épissage par le biais d'un variant, nous avons développé un nouveau score pour détecter l'apparition de tels sites d'épissage. Pour construire ce score, nous avons testés l'association des scores MES, SSF et ESRs. En ce qui concerne les ESRs, il a été utilisé le score ESRs défini par Ke et collaborateurs [24]. Afin de s'affranchir d'un problème de sur-ajustement, nous n'avons pas utilisé les données issues des 2 784 variants pour développer ce score. En effet, nous avons extrait l'ensemble des sites d'épissage des transcrits présents dans la base de données Ensembl, les données négatives correspondant aux positions des motifs AG/GT présents dans la séquence de ces transcrits, en dehors des sites consensus. Ainsi 202 989 656 sites, dont 530 931 sites d'épissage utilisés, ont servi à développer ce score. Deux-tiers de ce jeu de données (135 326 351 sites) ont été utilisés pour l'apprentissage, et le tiers restant (67 663 176 sites) pour la validation et la définition d'un seuil optimal de détection. Ainsi pour la recherche de création de site d'épissage par un variant, il est d'abord sélectionné dans l'environnement de ce variant, les possibles sites d'épissages (AG/GT) dont le score est renforcé par rapport à la séquence sauvage (Figure 49). Le score de ces possibles sites d'épissage est comparé au seuil de détection précédemment défini sur les données d'Ensembl, afin de conclure ou non sur la création d'un site d'épissage.

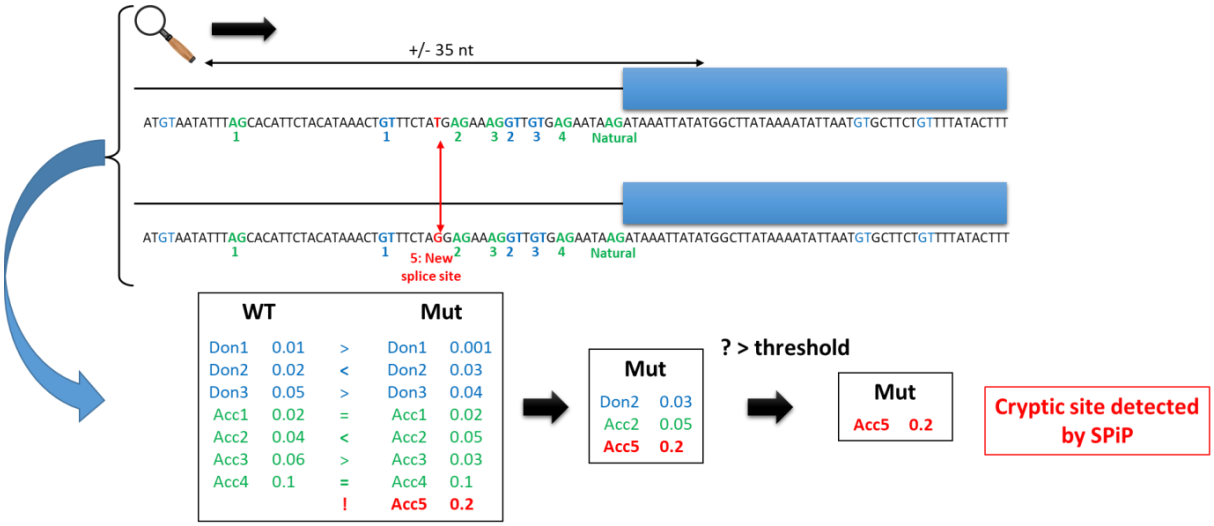


Figure 49 : La stratégie utilisée pour détecter une création d'un site d'épissage par un variant. Dans la zone de recherche, l'outil détecte 4 signaux AG (en vert) et 3 signaux GT (en bleu) dans la séquence sauvage et mutée, plus un cinquième AG (en rouge) dans la séquence mutée (*i.e.* site potentiel d'épissage *de novo*). L'outil compare le score de chaque site potentiel. Seul le second site donneur et le second site accepteur présentent un renforcement de score ainsi que le nouveau site accepteur (Acc5). Sur ces 3 sites, seulement Acc5 possède un score supérieur au seuil de détection et est donc prédit comme une création d'un nouveau site d'épissage.

Grâce à l'introduction de MES pour le tract polypirimidique et d'un nouveau score pour la détection des nouveau site d'épissage, SPiP offre ainsi la possibilité de prédire les modifications des motifs d'épissage par un variant quel que soit sa position (Figure 50). Cependant pour pouvoir conclure sur la splicéogénicité d'un variant, il a nécessité des seuils d'interprétation des différents scores de prédiction. En ce qui concerne SPiCE, BPP et la détection de nouveau site d'épissage, ces règles ont été précédemment définies [260], [262]. A propos de MES, il a été appliqué les recommandations émises par le groupe GGC, préconisant un seuil de 15 % [254]. Pour l'outil ΔtESRseq, il n'existait pas de recommandation pour l'interprétation de ce score. Aussi nous avons défini un seuil décisionnel sur la base des données issues des 2 784 variants. Pour cela nous avons extrait de cette collection les variants exoniques hors des sites consensus et induisant soit un saut d'exon ou sans effet sur l'épissage. Au total 1 068 variants exoniques ont été utilisés. Afin de réduire le risque de sur-ajustement nous avons aléatoirement sélectionné 100 variants sur ces 1 068 variants. Puis sur ce jeu aléatoire, nous avons défini un seuil optimal (score de Youden). Ces deux dernières étapes ont été répétées 1 000 fois. Sur les 1 000 seuils optimaux obtenues, nous avons pris le seuil médian comme seuil décisionnel. De cette façon, nous avons obtenus le seuil de -1.10 comme seuil décisionnel de l'outil ΔtESRseq.

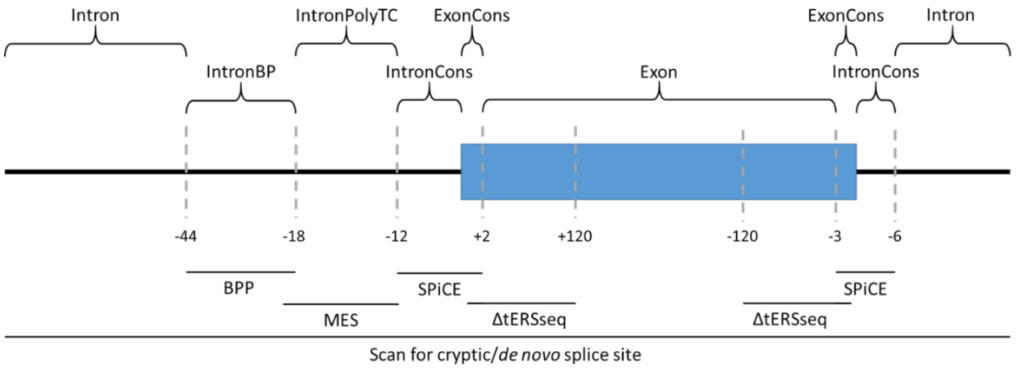


Figure 50 : Les outils de prédiction utilisés pour chaque motif d'épissage. IntronBP : Branch point area, IntronPolyTC : polypyrimidine tract, IntronCons : Intronic consensus splice site, ExonCons : Exonic consensus splice site.

Une fois l'ensemble des règles décisionnelles définies pour SPiP, nous avons testé l'efficacité de SPiP pour prédire un variant splicéogénique à partir des 2 784 variants. SPiP a atteint une exactitude de 80.76 %, une sensibilité de 90.96 % et une spécificité de 70.87 %. Au regard des défauts d'épissage, nouveau site d'épissage, saut d'exon, pseudo-exon et intron rétention, ces évènements étaient détectés avec une sensibilité de respectivement 96.03 % ; 89.77 % ; 82.96 % et 81.25 %. Puis nous avons comparé SPiP à SPANR [173] et à l'outil SpliceAI récemment publié [298]. Ces deux derniers outils intègrent également plusieurs motifs d'épissage pour prédire un variant splicéogénique. SPANR atteint une sensibilité et une spécificité de respectivement 78.37 % et 72.37 %. SpliceAI a montré une grande spécificité de 98.26 % (contre 70.87 % pour SPiP). Cependant SpliceAI a une moindre maitrise du risque de faut négatif avec une sensibilité de 70.71 % (contre 90.96 % pour SPiP).

Afin d'attribuer à chaque prédiction de SPiP une probabilité d'altération de l'épissage en fonction de la localisation du variant, nous avons procédé en trois étapes.

- Estimation de la proportion de variants splicéogéniques indépendamment des prédictions de SPiP et de leurs localisations dans le gène.
- Création d'un jeu de données équilibré, c'est-à-dire dont la proportion de variant splicéogéniques est en accord avec celle précédemment estimée, en intégrant des variants supposés sans effet au niveau de l'épissage.
- A partir de ce nouveau jeu de données, détermination des probabilités d'impacter l'épissage en fonction des prédictions fournies par SPiP et de la localisation du variant.

Pour estimer la proportion de variants splicéogéniques indépendamment des prédictions de SPiP et de leur localisation dans le gène, nous avons utilisé une approche bayésienne. Tout d'abord il a été admis que parmi les 2 784 variants, la proportion de variants splicéogéniques au niveau exonique est représentative de la probabilité qu'un variant exonique impacte l'épissage. En effet, au sein des 2 784

variants, la proportion de variants splicéogéniques est de 26.88 % (414/1 540). Or les données de la littérature montrent une proportion de 23.5 % (min : 10.3 % ; max : 48.0 %) [48], [49, p.], [99], [299]–[301]. Le théorème de Bayes nous dit que la probabilité qu'un variant exonique impacte l'épissage ($P_{exon}(spliceogenic)$) est :

$$P_{exon}(spliceogenic) = \frac{P(spliceogenic) \times P_{spliceogenic}(exon)}{P(exon)}$$

Où $P(spliceogenic)$ est la probabilité qu'un variant soit splicéogénique indépendamment de sa position dans le gène. $P_{spliceogenic}(exon)$ est la proportion de variants splicéogéniques situés dans un exon. $P(exon)$ est la probabilité qu'un variant soit situé dans un exon. De là nous pouvons en déduire que la probabilité qu'un variant soit splicéogénique quel que soit sa position est :

$$P(spliceogenic) = \frac{P(exon) \times P_{exon}(spliceogenic)}{P_{spliceogenic}(exon)}$$

Pour estimer $P(exon)$, nous avons utilisé les données du projet 1000 Genomes. Ce dernier rapportait plus de 80 millions de positions mutées du génome humain. Puis pour $P_{spliceogenic}(exon)$, nous avons exploité les données des 1 294 variants splicéogéniques rapportés parmi les 2 784 variants. Ainsi nous avons déterminé que la proportion de variants splicéogéniques indépendamment de sa position dans le gène est de 2.69 % ($CI_{95\%}$ [2.57 % – 2.82 %]).

Parmi les 2 784 variants, nous avons rapporté une proportion de variants splicéogéniques de 43 %. De ce fait, pour créer un jeu de données en accord avec l'estimation de la proportion de variants splicéogéniques, nous avons ajouté aux 2 784 variants un ensemble de variants présumés sans impact sur l'épissage. Ces variants présumés neutres ont été définis comme les variants ayant une MAF > 5% pour les gènes rapportés dans notre jeu de données, extrait *via* l'UCSC *genome browser*. Ainsi, 45 000 variants, parmi les 55 487 variants remplissant ces critères, ont été aléatoirement sélectionnés. Puis ces variants ont été intégrés aux 2 784 variants. Le jeu de données final contient donc 47 784 variants avec une proportion de variants splicéogéniques de 2.7 %.

Grâce à ce nouveau jeu de données de 47 784 variants, nous avons pu estimer les probabilités qu'un variant impacte l'épissage en fonction des prédictions de SPiP et de la localisation du variant. Si SPiP prédit une altération, alors ces probabilités d'altération de l'épissage correspondent aux VPP de l'outil. A l'inverse pour les prédictions négatives, ces probabilités d'altération de l'épissage correspondent aux 1- VPN. Pour illustrer ces VPP et VPN, nous avons représenté de manière schématique ces valeurs en fonction de la position du variant (Figure 51). Les VPP les plus faibles correspondent à la détection de nouveau site d'épissage (cryptique et *de novo*) étant donné le nombre conséquent de variants pouvant créer un nouveau site. Par contre, SPiP montre des VPP optimales pour les séquences consensus, le tract polypyrimidique, le point de branchement et pour les séquences ESRs. Par ailleurs les VPN de SPiP

sont toutes supérieures à 90 % [254], [260], [302]–[304]. A noter que ces probabilités ont été calculées à partir de classes de variants définies selon les prédictions de SPiP et la localisation sur le transcrit. Elles ne sont pas utilisées pour un réajustement *a posteriori* des prédictions de SPiP, ce qui conduirait nécessairement à un sur-ajustement de l'outil.

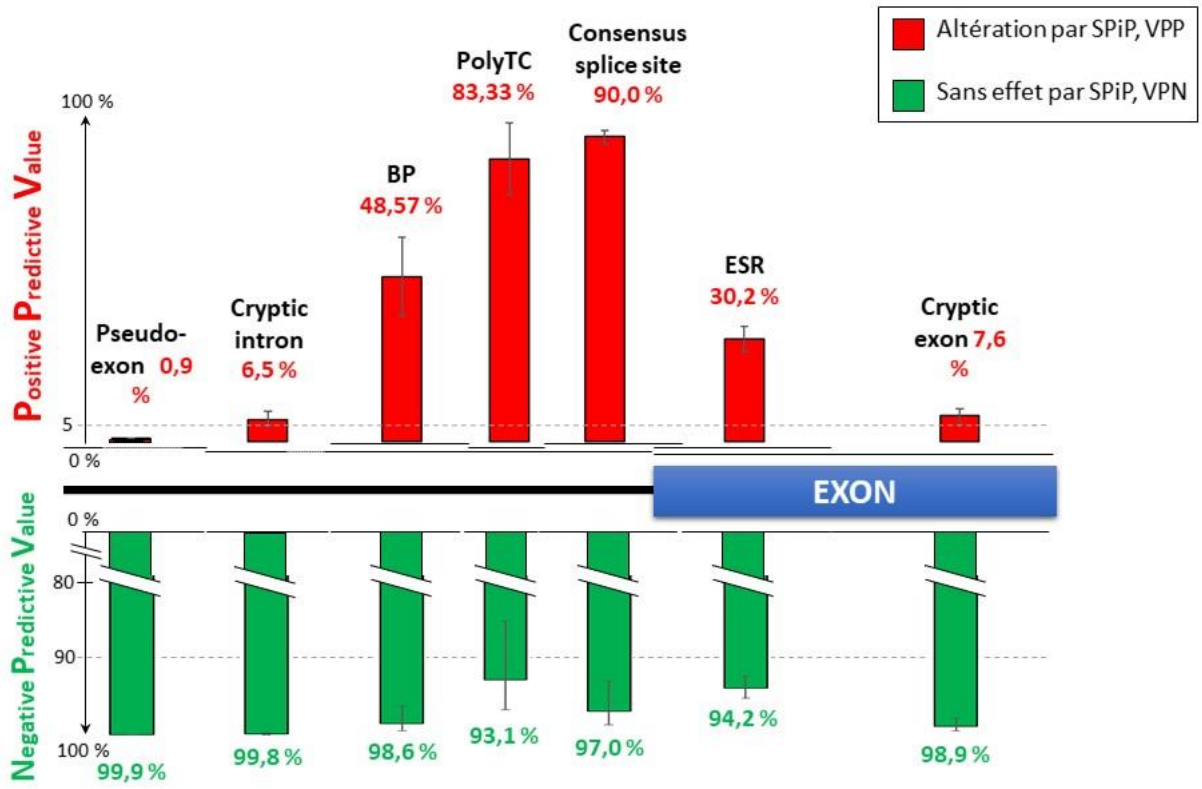


Figure 51 : Représentation simplifiée des VPP (valeur prédictive positive) et VPN (valeur prédictive négative) de SPiP en fonction de la position d'un variant.

SPiP se révèle donc être un outil de décision à large échelle pour aider à identifier des variants splicéogéniques parmi le déluge de variants identifiés par NGS. Nous proposons SPiP sous forme d'un logiciel interfacé accessible à https://sourceforge.net/projects/splicing-prediction-pipeline/ et d'une version Linux accessible à https://github.com/raphaelleman/SPiP.

IV. SpliceLauncher, un outil pour la détection, l'annotation et la quantification des jonctions alternatives à partir de données de RNA-seq : Article III

Le présent travail est une *application note* acceptée pour publication par *Bioinformatics*. Les données supplémentaires de cet article sont en **ANNEXE D**.

Etant donné la volumétrie des données RNA-seq, l'utilisation d'outils bioinformatiques et biostatistiques pour analyser ces données est indispensable. S'il existe de nombreux outils pour l'étude différentielle de l'expression des gènes, il n'en est pas de même pour l'étude des jonctions d'épissage. Aussi nous avons développé pour ce travail un nouvel outil nommé SpliceLauncher dédié aux jonctions d'épissage.

SpliceLauncher intègre un pipeline bioinformatique afin de détecter la présence de jonctions d'épissage à partir des fichiers FastQ. SpliceLauncher fournit une annotation humainement compréhensible de ces jonctions d'épissage en fournissant d'une part les coordonnées transcriptomiques, la nature de la jonction (saut d'exon, site alternatif, …) et d'autre part la nomenclature associée à cette jonction. Cette nomenclature a été définie au sein du consortium ENIGMA [264]. Puis, SpliceLauncher procède au calcul de l'expression relative des jonctions. Pour l'expression relative, SpliceLauncher propose une analyse statistique pour détecter les jonctions anormalement exprimées. Par ailleurs, lors de l'analyse, SpliceLauncher peut générer des fichiers PDF et BED pour visualiser les jonctions alternatives.

Ainsi, SpliceLauncher a permis d'identifier des profils d'épissage aberrants imputables à un variant et d'établir finement un panel d'épissages alternatifs utilisé dans le travail de Lopez-Perolio *et al.* pour le gène *PALB2* [264]. Il est également capable de traiter les données volumineuses générées par RNA-seq *whole transcriptome*.

SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data.

Raphaël Leman[1,2,3], Valentin Harter[4], Alexandre Atkinson[1], Grégoire Davy[1,2,3], Antoine Rousselin[1], Etienne Muller[1], Laurent Castéra[1,2,3], Fréderic Lemoine[5,6], Pierre de la Grange[7], Marine Guillaud-Bataille[8], Dominique Vaur[1,2,3], Sophie Krieger[1,2,3]
[1]Laboratoire de biologie et de génétique du cancer, Centre François Baclesse, 14076 Caen, France, [2]Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders, Normandie Univ, UNIROUEN, Normandy Centre for Genomic and Personalized Medicine, 76031 Rouen, France, [3]Normandie Univ, UNICAEN, 14000 Caen, France, [4]Cancéropôle Nord-Ouest Data Processing Centre, CLCC François Baclesse, 14076 Caen, France, [5]Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France, [6]Hub Bioinformatique et Biostatistique, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France, [7]GenoSplice technology, Paris, France, [8]Gustave Roussy, Université Paris-Saclay ,Département de Biopathologie, 94805 Villejuif, France.

1. Abstract

Summary: Alternative splicing is an important biological process widely analyzed in molecular diagnostic settings. Indeed, a variant can be pathogenic by splicing alteration and a suspected pathogenic variant (*e.g.* truncating variant) can be rescued by splicing. In this context, detecting and quantifying alternative splicing is challenging. We developed SpliceLauncher, a fast and easy to use open source tool that aims at detecting, annotating and quantifying alternative splice junctions at high resolution.
Availability: SpliceLauncher is available at https://github.com/raphaelleman/SpliceLauncher.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

2. Introduction

In the next-generation sequencing era, transcriptome analysis is a major challenge as more than 90 % of genes present an alternative splicing [30]. Many tools are already available to study differential gene expression (reviewed by [140]) and to quantify isoforms (reviewed by [305]). However, there are no tools to annotate finely alternative junctions. Developing such a tool will permit the researcher to determine the profile of alternative splicing for one or several genes in a clinical context. Knowledge of the alternative splicing allows to discover new transcripts and ultimately helps geneticists to interpret the impact of variants at the RNA level in a molecular diagnostic setting, as recently highlighted in [264]. Therefore, we propose a new bioinformatics tool, SpliceLauncher, that focuses on the detection, annotation and quantification of alternative junctions.

3. Methods

SpliceLauncher, with its corresponding RNAseq pipeline, aims at easing the process of analyzing splice junctions from RNAseq data. The RNAseq pipeline (Fig. S1) computes a junction read count matrix (Figure 52A) from raw FastQ files with the help of STAR v2.6.0 [114], Samtools v1.6 [107] and Bedtools v2.24.0 [116]. SpliceLauncher then uses this junction read count matrix to annotate and detect abnormally expressed junctions.

The first step requires an annotation file (table S1) that contains the structure of the transcripts to align the junctions on transcripts (Figure 52B). An example of such annotation file, based on RefSeq database [6], is provided with SpliceLauncher sources. SpliceLauncher also provides a tool that assists the users in creating their own annotations.

In the second step, SpliceLauncher starts by annotating each junction with the name of the gene it belongs to. By default, all transcripts per gene are used as reference. However, the users can either define a transcripts list of interest or force SpliceLauncher to use one transcript by gene according to the number of junctions supporting this transcript (Figure 52C). The transcripts will be used as reference to determine alternative splicing events. To reduce background noise, SpliceLauncher removes duplicate junctions having the same genomics coordinates but on the opposite strand (Figure 52C) (useful with stranded libraries).

Then, each junction is annotated with the kind of splicing event it supports, *i.e.* natural junction, junction supporting an exon skipping, an alternative 5' splice site or an alternative 3' splice site (Figure 52D). All junctions whose mapping coordinates fall outside the transcript are annotated as "Outside Transcript". The annotation "Event too complex" refers to junctions whose start and end coordinates do not correspond to any known transcripts. Moreover, SpliceLauncher assigns to each junction transcriptomic coordinates as well as a unique and well characterized name as recommend by López-Perolio et al., 2019 (Fig. S2). For example, the junction chr13:32,915,333-32,920,963 becomes Δ12 of *BRCA2* (NM_000059).

SpliceLauncher performs a calculation of relative expression for each annotated junctions. The main principle of this calculation is to derive a ratio (expressed as a percentage), of the alternative junction read count over the read count of the relevant physiological junction (Fig. S3). From this relative expression, SpliceLauncher is able to detect abnormal junction expression (Figure 52D). To do so, this analysis models the distribution of the relative expression for each junction, either as a gamma distribution or a negative binomial distribution. Fitting the distribution to the data during this step excludes outlier expression values, as determined by the Tukey method. For each distribution, SpliceLauncher calculates the probability that a sample shows abnormal expression of this junction [131].

At the end of analysis, SpliceLauncher generates a report (table S2), a graphic representation of alternative splicing (Figure 52E and Fig. S4) and a BED annotation file to visualize junctions in a genome browser such as UCSC Genome browser (Fig. S5 and table S3). This permit visualization of the alternative splicing in a dynamic and user-friendly way.

4. Use case

We studied the splicing patterns of genes involved in the predisposition of breast and ovarian cancers by targeted short-read RNAseq on leucocytes and breast tissue samples of healthy donors or carriers of splicing variants. First, our work contributed to defined a comprehensive characterization of alternative splicing in *PALB2* [264]. Alterations of *PALB2* gene are closely associated with predisposition of breast cancer and the knowledge of alternative splicing helps for classification of nucleotidic variants. Eighty-one alternative events were newly described, SpliceLauncher permitted to characterize 91.4% of these events. Second, the statistical analysis of SpliceLauncher was assayed on samples with the variant c.4096+3A>T of *BRCA1* (NM_007294), inducing Δ11q(3309) (shift of donor splice site) and Δ11 (exon 11 skipping), confirmed by RT-PCR (Fig. S7A). We used two biological replicates of lymphoblastoid cell lines carriers and eight controls. SpliceLauncher detected 17,940 junctions and 1,530 junctions were adjusted to Negative binomial or Gamma distribution. Six junctions were detected with abnormal expression in the variant carrier samples, with only two junctions in *BRCA1* gene (data not shown). The junction Δ11q(3309) was expressed at 136.8 % and 119.8 % in variant carrier samples instead of 5.35 %, the average expression in controls sample (Fig. S7B). SpliceLauncher calculated a p-value below 0.001. The junction Δ11 was expressed at 4.6 % and 8.5 % in variant carrier samples instead of 0.13 %, the average expression in controls sample. SpliceLauncher calculated a p-value of 0.045 and 8.12×10^{-6}. SpliceLauncher was also successfully tested on whole transcriptome RNAseq with the published data of [306], and detected 411,285 junctions. Among these junctions, 75.4 % were annotated (310,316/411,285). From 34,227 transcripts, SpliceLauncher selected 21,388 transcripts as having one transcript by gene (table S4).

5. Conclusion

SpliceLauncher provides a relevant tool to determine the alternative splicing from targeted or whole transcriptome RNAseq data. The tool can deal with a great number of junctions and displays graphical results which facilitates the interpretation of alternative splicing.

6. Acknowledgements

We thank Alexandra Martins and Pascaline Gaildrat to initiate this project. We acknowledge Raphaël Lanos and Logan Walker for their critical reading. We also thank Germain Paimparay for his bioinformatics counseling.

Conflict of Interest: none declared.

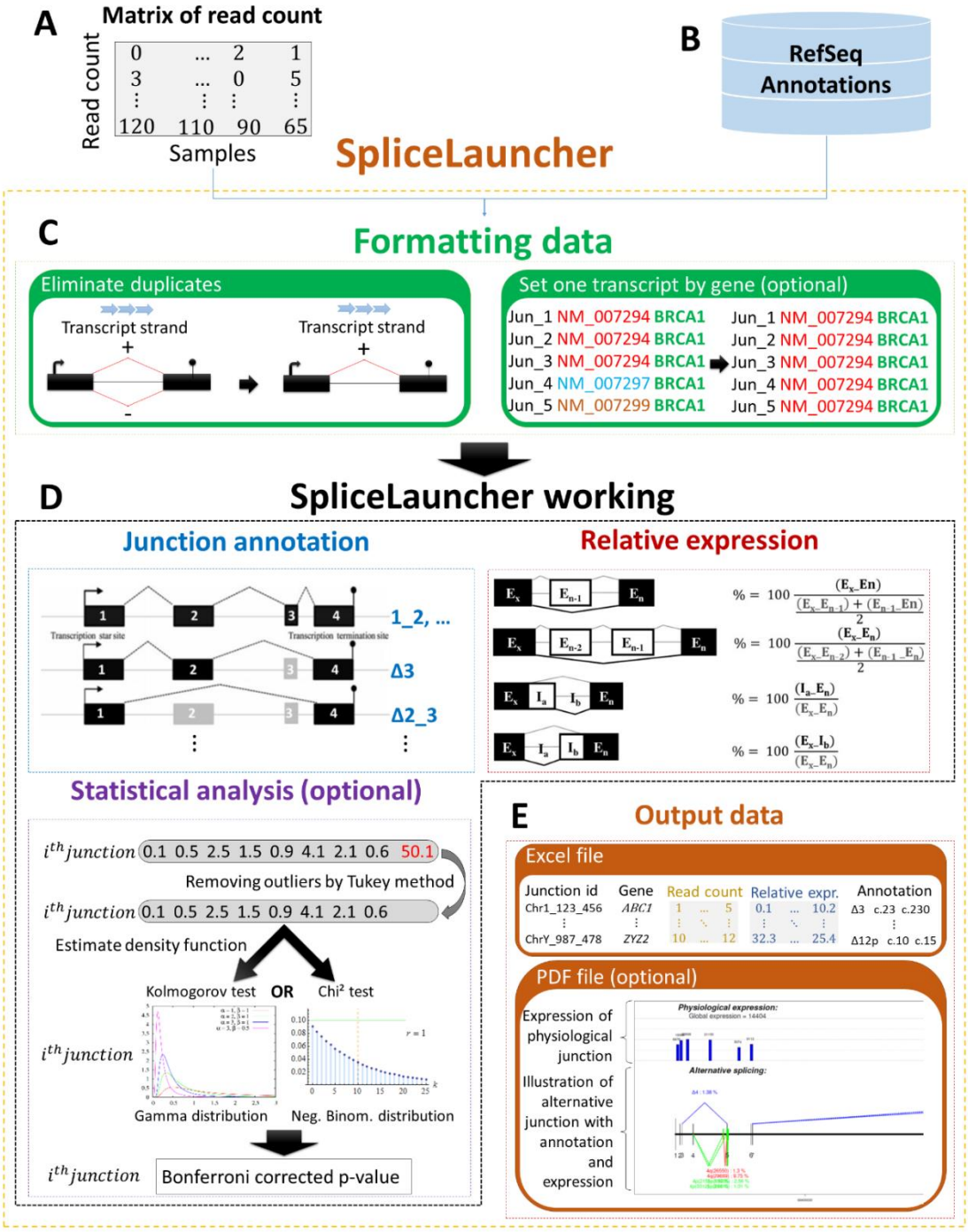


Figure 52 : SpliceLauncher analysis pipeline. **A**. the junctions read count matrix, **B**. the list of transcripts contained in the RefSeq database. **C**. SpliceLauncher removes irrelevant junction and sets one transcript per gene. **D**. SpliceLauncher determines the alternative events, calculates the relative expression over natural junction and detects the abnormally expressed junctions. **E**. SpliceLauncher provided an excel file or a tabulated-text file (optional) with the raw data and the illustration of the alternative junctions.

V. L'impact de l'épissage alternative dans la classification des variant *PALB2* selon les recommandations de l'ACMG-AMP 2015, un rapport ENIGMA : article N°IV

Le présent travail a été publié dans *Journal of Medical Genetics* en 2019 (doi : http://dx.doi.org/10.1136/jmedgenet-2018-105834). Les données supplémentaires de cet article sont en **ANNEXE E**, les tableaux contenant les épissages alternatifs *high-confidence* et *low-confidence* ainsi que la liste des amorces utilisées (*tables S1 to S3*) sont disponibles en ligne à https://jmg.bmj.com/content/56/7/453.

Les variants délétères de *PALB2* sont associés au syndrome HBOC avec un risque similaire aux variants délétères de *BRCA2*. Ainsi, le gène *PALB2* a récemment été intégré au panel de gènes testé pour le diagnostic moléculaire du syndrome HBOC. Ne disposant pas d'un recul aussi important que pour *BRCA1*/*BRCA2*, le modèle multifactoriel n'est pas applicable pour déterminer la pathogénicité des variants. Aussi, il est préférable d'utiliser la classification catégorielle proposée par l'ACMG-AMP. Les variants tronquants (non-sens ou hors phase) ainsi que les variants des sites canoniques (AG/GT) sont considérés comme à haut risque d'être pathogènes, soit PVS1 dans les catégories de l'ACMG-AMP. Or l'ACMG-AMP met en garde contre le risque de sauvegarde de la fonctionnalité des protéines malgré la présence de ces variants PVS1 du fait, entre autre, de l'existence d'épissages alternatifs. De même qu'une expression tissu spécifique des transcrits alternatifs peut, selon l'ACMG, modifier l'interprétation de l'effet de variants. Aussi, pour cette étude nous avons caractérisé les transcrits alternatifs du gène *PALB2* afin d'évaluer s'ils étaient soumis à une expression tissu-spécifique et si ces événements d'épissage pouvaient impacter l'interprétation des variants PVS1.

Des données de RNA-seq et d'électrophorèse capillaire de produits de RT-PCR de plus de 100 échantillons issus de différentes sources biologiques ont été générées et pré-analysées par différents laboratoires puis centralisées. Les épissages alternatifs ont été caractérisés en les séparant en deux groupes *high-confidence* et *low-confidence* selon le niveau de preuve apporté par les données. Ces événements d'épissage ont été annotés en fonction de leur impact probable sur la protéine. Nous avons aussi testé *in vitro* l'impact de sept variants abolissant les sites naturels d'épissage du pré-ARNm du gène *PALB2* (c.212-1G>A, c.1684+1G>A, c.2748+2T>G, c.3113+5G>A, c.3350+1G>A, c.3350+4A>C, and c.3350+5G>A). Nous avons extrapolé ces résultats, afin de déterminer parmi les variants situés au niveau des sites canoniques, lesquels peuvent être considérés PVS1 et ceux pour lesquels la classe PVS1 n'est pas garantie en l'absence d'études complémentaires.

Nous avons ainsi identifié 88 évènements d'épissage dont 44 étaient des évènements *high-confidence* et retrouvés dans les différents tissus biologiques. Ceci a permis d'exclure une expression tissu-spécifique des transcrits de *PALB2*. Parmi ces évènements d'épissage, 15 préservent le cadre de lecture et parmi eux seulement 6 évènements au niveau des sites accepteurs pourraient produire une protéine fonctionnelle. Ainsi, la catégorie PVS1 est garantie pour tous variants dans les sites canoniques donneurs de *PALB2*. Cependant, pour les sites accepteurs des exons 2, 5, 7 et 10, la catégorie PVS1 n'est pas

garantie. Par conséquent, des variants situés dans ces sites accepteurs d'épissage ne peuvent être considéré de suite comme pathogènes ou probablement pathogènes sans données fonctionnelles et cliniques supplémentaires.

Alternative Splicing and ACMG-AMP-2015 Based Classification of *PALB2* Genetic Variants: an ENIGMA Report

Irene López-Perolio, PhD[1,†], Raphaël Leman, PharmD[2,†], Raquel Behar, MSc[1], Vanessa Lattimore, PhD[3], John F. Pearson, PhD[3], Laurent Castéra, PharmD-PhD[2], Alexandra Martins, PhD[4], Dominique Vaur, PharD[2], Nicolas Goardon, PhD[2], Grégoire Davy, PharmD-PhD[2], Pilar Garre, PhD[1], Vanesa García Barberán, PhD[1], Patricia Llovet, PhD[1], Pedro Pérez-Segura, MD, PhD[1], Eduardo-Díaz Rubio, MD, PhD[1], Trinidad Caldés, PhD[1], Kathleen S. Hruska[5], PhD, Vickie Hsuan, MSc[6], Sitao Wu, PhD[6], Tina Pesaran, MSc, CGC[6], Rachid Karam, MD, PhD[6], Johan Vallon-Christersson, PhD[7], Åke Borg, PhD[7], kConFaB Investigators[8,9], Alberto Valenzuela-Palomo,Bsc[10], Eladio A. Velasco, PhD[10], Melissa Southey. PhD[11], Maaike P.G. Vreeswijk, PhD[12], Peter Devilee, PhD[12], Anders Kvist, PhD[7], Amanda B. Spurdle, PhD[13], Logan Walker, PhD[3], Sophie Krieger, PharmD-PhD[2,#], Miguel de la Hoya, PhD[1,#].
[†]contributed equally to this paper, [#]contributed equally to this paper.
[1]Molecular Oncology Laboratory CIBERONC, Hospital Clinico San Carlos, IdISSC (Instituto de Investigación Sanitaria del Hospital Clínico San Carlos), Madrid, Spain; [2]Laboratory of Clinical Biology and Oncology, Center François Baclesse, Inserm U1245 Genomics and Personalized Medicine in Cancer and Neurological Disorders, Normandy University Caen, France.[3]Department of Pathology and Biomedical Science. University of Otago Christchurch. New Zealand; [4]Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders, UNIROUEN, Normandie Université, Normandy Centre for Genomic and Personalized Medicine, Rouen, France; [5]GeneDx, Gaithersburg, Maryland, USA; [6]Ambry Genetics, Aliso Viejo, CA, USA; [7]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Sweden; [8]Peter MacCallum Cancer Centre, Grattan Street, Melbourne, VIC, Australia; [9]The Sir Peter MacCallum Department of Oncology University of Melbourne, Parkville, Australia; [10]Splicing and genetic susceptibility to cancer, Instituto de Biología y Genética Molecular (CSIC-UVa), Valladolid, Spain;[11]Genetic Epidemiology Laboratory, Department of Clinical Pathology, The University of Melbourne, Melbourne, VIC, Australia;[12]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [13]Molecular Cancer Epidemiology Laboratory. QIMR Berghofer Medical Research Institute. Brisbane. Australia.

Corresponding author

Miguel de la Hoya, PhD, Molecular Oncology Laboratory, Hospital Clínico San Carlos, Madrid, 28040, Spain. Tel:+34913303348; Fax:+34913303544; e-mail: mdhoya@hotmail.com

1. Abstract

Background: *PALB2* monoallelic *loss-of-function* germ-line variants confer a breast cancer risk comparable to the average *BRCA2* pathogenic variant. Recommendations for risk reduction strategies in carriers are similar. Elaborating robust criteria to identify *loss-of-function* variants in *PALB2*-without incurring overprediction- is thus of paramount clinical relevance. Towards this aim, we have performed a comprehensive characterization of alternative splicing in *PALB2*, analyzing its relevance for the classification of truncating and splice site variants according to the 2015 American College of Medical Genetics and Genomics-Association for Molecular Pathology guidelines.

Methods: Alternative splicing was characterized in RNAs extracted from blood, breast and *fimbriae*/ovary related human specimens (N=112). RNA-seq, RT-PCR/CE, and CloneSeq experiments were performed by 5 contributing laboratories. Centralized revision/curation was performed to assure high-quality annotations. Additional splicing analyses were performed in *PALB2* c.212-1G>A,

c.1684+1G>A, c.2748+2T>G, c.3113+5G>A, c.3350+1G>A, c.3350+4A>C, and c.3350+5G>A carriers. The impact of the findings on PVS1 status was evaluated for truncating and splice site variant.
Results: We identified 88 naturally occurring alternative splicing events (81 newly described), including 4 in-frame events predicted relevant to evaluate PVS1 status of splice site variants. We did not identify tissue-specific alternate gene transcripts in breast or ovarian related samples, supporting the clinical relevance of blood-based splicing studies.
Conclusions: PVS1 is not necessarily warranted for splice site variants targeting four *PALB2* acceptor sites (exons 2, 5, 7, and 10). As a result, rare variants at these splice sites cannot be assumed *pathogenic*/*likely pathogenic* without further evidences. Our study puts a warning in up to five *PALB2* genetic variants that are currently reported as *pathogenic/likely pathogenic* in ClinVar.
Keywords: *PALB2*, Splicing, Variant Classification, ACMG-AMP guidelines, PVS1

2. Introduction

Monoallelic *loss-of-function* (LoF) germ-line variants in *PALB2* predispose to breast cancer, with estimated absolute risks by age 80 ranging from 33% to 58%, depending on the family history [214], [307]. Excess risk for other cancers, such as pancreas, prostate, ovarian, and male breast cancer, is still under investigation. Currently, gene panel testing for breast cancer predisposition includes *PALB2* [307], and LoF germ-line variants in this gene are considered actionable findings in many settings, with proposed actions ranging from increased surveillance to prophylactic surgery [308]–[310]. Accordingly, classifying *PALB2* LoF variants is of paramount clinical relevance. Yet, the task is not trivial, as proved by the large number of variants of uncertain significance (VUS) still existing in genes that have been extensively studied, such as *BRCA1* or *BRCA2* [311].

In the research setting, truncating (nonsense or frame-shift) variants predicted to induce Nonsense Mediated Decay (PTC-NMD variants) and canonical ±1,2 splice site variants (hereafter named splice site variants) at cancer predisposition genes are often assumed pathogenic/likely pathogenic LoF variants [214], [307]. However, in the clinical setting a more conservative approach is recommended. According to *the American College of Medical Genetics and Genomics-Association for Molecular Pathology* (ACMG-AMP) interpretation guidelines [232], a PTC-NMD or splice site variant is a very strong evidence of pathogenicity (PVS1), but not sufficient to classify the variant as pathogenic/likely pathogenic. Additional combinations of strong (PS), and/or moderate (PM), and/or supporting (PP) evidence of pathogenicity are required. Further, PVS1 is not warranted for every PTC-NMD/splice site variant. Indeed, the ACMG-AMP-2015 guidelines specify several caveats, including the possibility of: (i) *rescue* transcripts (alternate gene transcripts that skip the truncating variant, encoding functional or partially functional proteins, and resulting in reduced or no haplo-insufficiency), (ii) splice site variants producing transcripts with in-frame deletions/insertions retaining some or all functional capacity, and (iii) tissue-specific alternate gene transcripts [232]. Therefore, the accurate interpretation of *PALB2*

PTC-NMD and splice site variants according to the ACMG-AMP-2015 guidelines requires reliable information on both protein structure/function and alternative splicing.

To be more precise, *PALB2* PTC-NMD/splice site variants without direct risk estimates and/or functional data (a common scenario in genetic testing) should be classified as likely pathogenic only if PVS1 is warranted. For PTC-NMD variants, PVS1 is warranted if no *rescue* transcripts are predicted. For splice site variants the analysis is more complex. In addition to *rescue* transcripts, the possibility of the variant allele producing transcripts with in-frame alterations retaining coding potential should be considered, albeit predicting the precise nature of the transcripts produced by a splice site variant is challenging.

In recent years, the Evidence-based Network for the Interpretation of Germ-line Mutant Alleles (ENIGMA consortium) has conducted a comprehensive characterization of naturally occurring alternate gene transcripts in *BRCA1* and *BRCA2* [190], [191], exploring the impact of the findings for the clinical classification of genetic variants at the two loci. Major achievements were the identification of a subset of splice sites variants for which PVS1 was not necessarily warranted, the posterior demonstration that at least one allele containing a splice site variant, *BRCA1* c.[594-2A>C; 641A>G], does not increase breast cancer risk, and the observation that splicing assays may lead to erroneous clinical conclusions if alternate gene transcripts are not properly addressed [190], [191], [255], [312]. Recommendations based on these studies are documented in the *ENIGMA BRCA1/2 Gene Variant Classification Criteria* (https://enigmaconsortium.org) that support *BRCA1* and *BRCA2* expert panel review interpretation at ClinVar.

A recent study has identified alternate gene transcripts at the *PALB2* locus, but no inferences in relation to the clinical interpretation of genetic variants were made [131]. Here we undertake a comprehensive characterization of *PALB2* alternative splicing, exploring the possible relevance of the findings for the clinical classification of PTC-NMD and splice site variants according to the ACMG-AMP-2015 guidelines.

3. Methods

a. Identification of alternative splicing events

To characterize alternative splicing at the *PALB2* locus, we analyzed RNAs isolated from *112* specimens, including lymphoblastic cell lines not treated with the NMD-inhibitor puromycin (LCLs-Puro, *N=68*), matched replicates treated with puromycin (LCL+Puro, *N=1*), stimulated leukocytes cultures not treated with puromycin (sLEU-Puro, *N=6*), matched replicates treated with puromycin (sLEU+Puro, *N=3*), RNA stabilized peripheral blood samples (PAXgene, QIAGEN, *N=7*; Tempus, Thermofisher, *N=10*), non-malignant breast tissue samples from unrelated women (Breast, *N=12*; 10 corresponding to women with a diagnosis of breast cancer, of which 9 are included in SCAN-B,

ClinicalTrials.gov identifier: NCT02306096; 2 corresponding to women without a diagnosis of breast cancer included in CASOHAR trial NTC02560818), a human mammary epithelial cell (HMEC, *N=1*, 2 technical replicas included in the analysis), commercially available RNA from non-malignant breast tissue (Clontech 636576, *N=1*), normal ovarian *fimbriae* tissue samples from prophylactic oophorectomies performed in post-menopausal women without cancer (Fimbriae, *N=2*), and one pool of 3 non-malignant ovarian tissues (Clontech 636555, *N=1*).

Experiments were performed independently in 5 ENIGMA laboratories (Figure 53). Most samples were analyzed by targeted RNAseq (N=72) in Laboratory 1 (Supplemental Tables 1 and 2). Other samples were analyzed by whole transcriptome RNAseq (N=13) in Laboratories 2 and 3 (Supplemental Tables 1 and 2), by capillary electrophoresis of RT-PCR products (RT-PCR/CE, N=22) in Laboratory 4 **(**Supplemental Tables 1, 2, and 3, Supplemental Figures 1A and 1B**)**, and by whole-gene CloneSeq splicing analysis (N=5) in Laboratory 5 **(**Supplemental Figure 1B**)**. We later performed a centralized revision/curation of the data, including the search for putative tissue-specific alternate gene transcripts. To this end, we pooled together all data produced in LCLs±Puro, sLEU±Puro, PAXgene and Tempus samples (hereafter referred collectively as BLOOD), all data produced in non-malignant breast tissues, HMEC, and Clontech 636576 (hereafter referred as BREAST), and all data produced in non-malignant ovarian *fimbriae* and Clontech 636555 (hereafter referred as OVARY). The overall workflow is summarized in Figure 53 (See Supplemental Material Section 1 for further details).

b. Annotation of alternative splicing events.

We described all alternative splicing events according to HGVS guidelines, using as a reference the Ensembl transcript ENST00000261584.8 (NCBI RefSeq NM_024675.3). For the sake of simplicity, we also identified most events with a code that combines the following symbols: Δ (skipping of reference exonic sequences), ▼ (inclusion of reference intronic sequences), E (exon), I (intron), p (acceptor shift), q (donor shift), AFE (alternative first exon), and IVS± (located at intervening sequence). When necessary, the exact number of nucleotides skipped (or retained) is indicated. Events were annotated as well according to the confidence of the finding (high-confidence vs. lower-confidence), predictions on coding potential (LoF vs. uncertain), and relative quantification (expression level relative to the corresponding reference transcript). See Supplemental Material Section 2 and Supplemental Figures 2-5 for further details.

c. Analysis of PVS1 status (warranted vs. not warranted) for every possible PTC-NMD and splice site variant at the PALB2 locus.

To decide if PVS1 is warranted we used predictions based on: (i) the identification of alternate gene transcripts in control samples, and (ii) RNA splicing assays performed previously in carriers of *PALB2* splice site variants (Supplemental Table 4), and (iii) novel RNA splicing assays (Supplemental Table 4, Supplemental Figures 6A, 6B and 6C). In brief, we consider PVS1 warranted for PTC-NMD variants

only if no plausible *rescue transcripts* have been detected. Similarly, we consider PVS1 warranted for splice site variants only if all predicted RNA product are bona fide LoF transcripts. To predict possible RNA products we used splicing assays performed in carriers of splice site variants (assuming that other *PALB2* splice site variants targeting the same splicing site will produce similar transcripts). If no splicing assay was available for a particular splice site, we based predictions on alternate gene transcripts, as previously done for *BRCA1* and *BRCA2* [191], [312]. Further details are shown in Supplemental Material Section 3 and Supplemental Table 4.

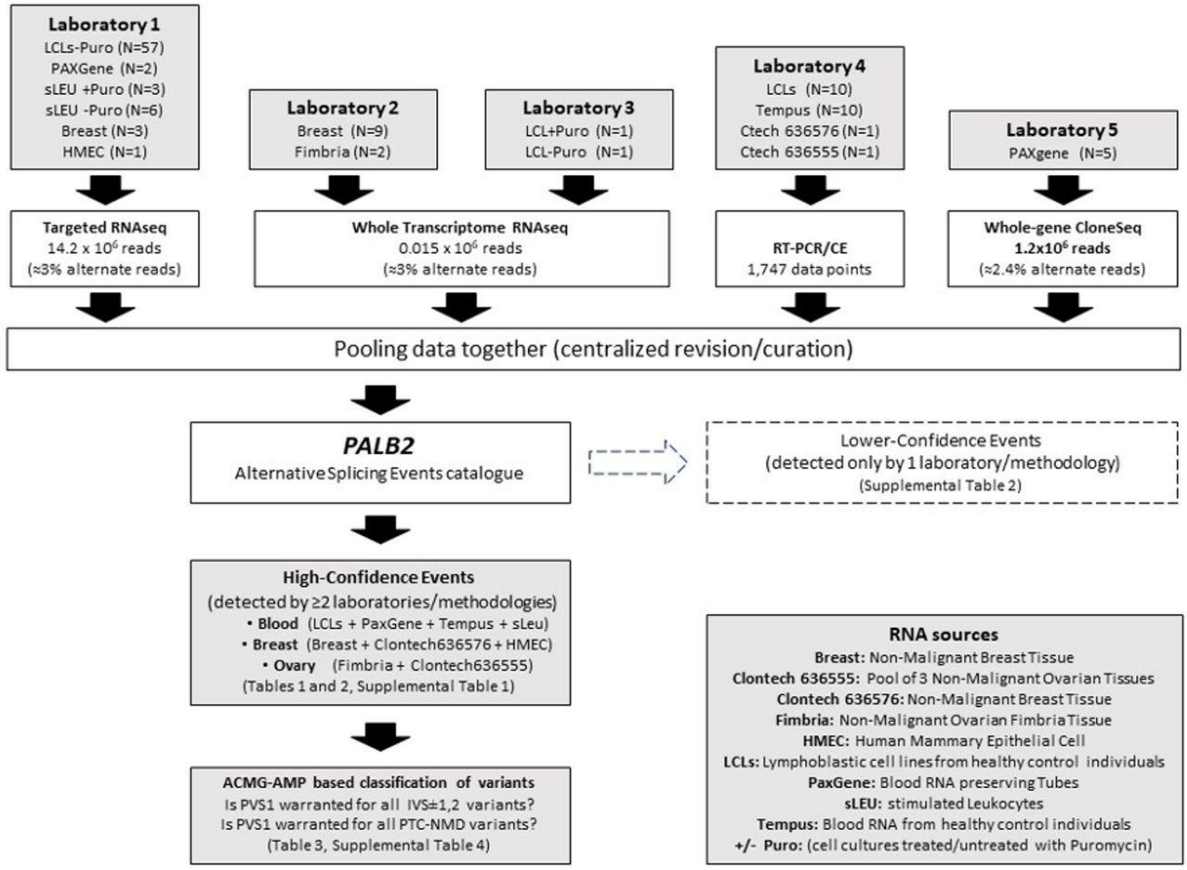


Figure 53 : Workflow. The workflow is followed by the Evidence-based Network for the Interpretation of Germ-line Mutant Alleles consortium to characterise the naturally occurring alternative splicing profile at the PALB2 locus in BLOOD-derived, BREAST-derived and OVARY-derived samples. RNAseq data were produced in five independent laboratories using different methodologies in unrelated samples. Laboratory 1 (Clinical Biology and Oncology Laboratory, Cancer Center François Baclesse, Normandy University Caen, France) performed targeted RNAseq analysis. Laboratories 2 (Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Sweden) and 3 (Department of Pathology and Biomedical Science, University of Otago Christchurch, New Zealand) performed whole transcriptome RNAseq. Laboratory 4 (Molecular Oncology Laboratory, Academic Hospital San Carlos, Madrid, Spain) performed capillary electrophoresis analysis of real-time PCR products (RT -PCR/CE). Laboratory 5 (Ambry Genetics) performed whole-gene CloneSeq alternative splicing analysis. As indicated, the overall contribution of targeted RNAseq reads to the analysis is roughly 1000× higher than that of whole transcriptome RNAseq. For instance, targeted RNAseq experiments end up with 13 754 118 reads aligned to reference exon-exon junctions, but only 459 186 reads supporting alternative splicing events (≈3%). The same percentage was observed in whole transcriptome RNA experiments, although the total number of reads was much lower (14 933 reads combining data from laboratories 2 and 3). RT - PCR/CE contributed 1747 data points (individual RT - PCR experiments performed with a particular combination of primers in individual samples, including technical replicas). CloneSeq analysis contributed 1.2×106 reads (≈2.4% of the reads supporting alternative splicing events). Data were pooled together, reviewed and cross-checked to end up with a list of high-confidence naturally occurring alternative splicing events (events detected by different techniques in different samples), and a list of lower-confidence splicing events (events not qualifying for higher confidence events). Finally, the possible relevance of high-confidence findings for the initial classification of canonical splicing site and PTC-NMD variants was explored. AC MG-AMP, American College of Medical Genetics and Genomics-Association for Molecular Pathology; HMEC, human mammary epithelial cell; LCL, lymphoblastic cell line; NMD, nonsense-mediated decay.

4. Results

We used RNA extracted from different human biological samples (blood-derived, breast and ovary; see 'Methods' section) to characterize naturally occurring alternative splicing at the *PALB2* locus. This study combined targeted RNAseq, whole-transcriptome RNAseq, RT-PCR/CE and whole-gene CloneSeq splicing analysis data that was independently produced at 5 contributing centers (Figure **53**). The analysis identified 44 naturally occurring alternative splicing events with high-confidence (**Supplemental Table 1**) and provided evidence for the existence of up to 44 additional (*lower-confidence* events, Supplemental Table 2 and Supplemental Material Section 2.2). Most events (37 out of 44 high-confidence and all *lower-confidence* events) have not been described previously in GENCODE (https://www.gencodegenes.org/) or the scientific literature to our knowledge.

Up to 15 high-confidence events preserved a bona fide open reading frame (i.e. a ORF spanning from the reference start codon to the reference termination codon, Table 10, protein column). Of these, nine were predicted to code for non-functional proteins, and the remaining six for proteins of uncertain functionality (Table 10, coding potential column). Twenty-nine high-confidence events did not preserve a bona fide ORF. All of them were predicted to code for non-functional proteins (Table 11).

Table 10 : High-confidence alternative splicing events at the PALB2 locus (In-frame events)

Designation[1]	Biotype[2]	RNA[3]	Protein[3]	Coding Potential[4]	Rationale[4]	BLOOD	BREAST	OVARY
▼(AFE600)+ Δ(E1)#	Terminal modification	r.1_28delins28+805_28+858	p.Asp2_Lys16delins17	Uncertain	Damaging to CC	yes	yes	-
▼(E1q9)	Donor shift	r.48_49ins48+1_48+9	(p.Lys16_Leu17ins3)	Uncertain	Uncertain impact on CC	yes	-	yes
Δ(E2p6)	Acceptor shift	r.49_54del	(p.Leu17_Lys18del)	Uncertain	Uncertain impact on CC	yes	yes	yes
Δ(E2)	Cassette	r.49_108del	(p.Leu17_Asn36del)	LoF	Damaging to CC	yes	-	yes
Δ(E4)	Cassette	r.212_1684del	(p.Glu71_Lys561del)	LoF	Damaging to ChAM	yes	yes	yes
Δ(E5p24)	Acceptor shift	r.1685_1708del	(p.Gly562_Lys569del)	Uncertain	No domain affected	yes	yes	yes
Δ(E6)¥	Cassette	r.2515_2586del	(pThr839_Lys862del)	LoF¥	Damaging to WD40¥	yes	yes	yes
▼(E7p42)	Acceptor shift	r.2586_2587ins2587-42_2587-1	(p.Lys862_Asn863ins14)	Uncertain	Uncertain impact on WD40	yes	yes	yes
Δ(E7)	Cassette	r.2587_2748del	(p.Arg863_Glu916del)	LoF	Damaging to WD40	yes	yes	yes
Δ(E9p30)	Acceptor shift	r.2835_2864del	(p.Ala946_Glu954del)	LoF	Damaging to WD40	yes	yes	yes
Δ(E9)	Cassette	r.2835_2996del	(p.Ala946_Gly1000del)	LoF	Damaging to WD40	yes	yes	yes
Δ(E9_E10)	Multi-cassette	r.2835_3113del	(p.Ala946_Trp1038del)	LoF	Damaging to WD40	yes	yes	-
Δ(E10p3)	Acceptor shift	r.2997_2999del	(p.Gly1000del)	Uncertain	Uncertain impact on WD40	yes	yes	-
Δ(E10)	Cassette	r.2997_3113del	(p.Gly1000_Trp1038del)	LoF	Damaging to WD40	yes	yes	yes
Δ(E11_E12)†	Multi-cassette	r.3114_3350del	(p.Asn1039_Arg1117del)	LoF	Damaging to WD40	yes	yes	yes

[1] See Supplemental Material Section 2.1 and Supplemental Figure 2 for details. [2] Biotype according to ENCODE [313]. [3]RNA and predicted protein described according to the Human Genome Variation Society guidelines at http://varnomen.hgvs.org/, using Ensembl transcript ENST00000261584.8 as a reference. [4]Uncertain coding potential if the transcript encodes a protein predicted to preserve (or partially preserve) functional capacity. See online supplemental material section 2.3 and figure 4 for further details. †Only Δ11_12 described previously in the literature [131]. [#]Only▼(AFE600)+Δ(E1) described in GENCODE (comprehensive gene annotation from GENCODE release 26 retrieved through Ensembl at http://www.ensembl.org/). [¥]Δ(E6) transcripts code for a hypomorphic protein (instable, but with residual activity)[314]. CC (N-terminal coiled-coil domain). ChAM (chromatin associated motif). WD40 (WD40 Δ-propeller C-terminal domain).

Table 11 : High-confidence alternative splicing events at the PALB2 locus (PTC-NMD events)

Designation[1]	Biotype[2]	RNA[3]	Protein	Coding Potential	BLOOD	BREAST	OVARY
Δ(E1q169)	donor shift	r.-121_48del	non-coding	LoF	yes	yes	yes
Δ(E1q17)†¶	donor shift	r.32_48del	p.Cys11Phefs*25	LoF	yes	yes	yes
▼(E1q337)	donor shift	r.48_49ins48+1_48+337	p.Leu17Valfs*19	LoF	yes	-	-
IVS1-463▼(134)†,¶	cassette	r.48_49ins49-463_49-330	p.Leu17Valfs*11	LoF	yes	yes	-
▼(E2p26)	acceptor shift	r.48_49ins49-26_49-1	p.Leu17Tyrfs*9	LoF	yes	-	yes
▼(I2)	Intron retention	r.108_109ins108+1_109-1	p.R37_S1186delins11	LoF	yes	-	-
▼(E3p36)	acceptor shift	r.108_109ins109-36_109-1	p.Arg37_Ser1186deldelins11	LoF	yes	yes	yes
▼(E4p25)	acceptor shift	r.211_212ins212-25_212-1	p.Glu71Valfs*10	LoF	yes	-	-
Δ(E4_E5)†,¶	multi-cassette	r.212_2514del	p.Glu71Aspfs*1	LoF	yes	yes	-
Δ(E5p139)	acceptor-shift	r.1685_1823del	p.Gly562Valfs*19	LoF	yes	yes	-
Δ(E5)	cassette	r.1685_2514del	p.Gly562Aspfs*1	LoF	yes	-	-
▼(E6p28)	acceptor shift	r.2514_2515ins2515-28_2515-1	p.Glu840Asnfs*9	LoF	yes	yes	yes
▼(E7p20)	acceptor shift	r.2586_2587ins2587-20_2587-1	p.Pro864Cysfs*13	LoF	yes	-	yes
Δ(E7p2)	acceptor shift	r.2587_2588del	p.Asn863Serfs*20	LoF	yes	yes	-
Δ(E7p10)	acceptor shift	r.2587_2596del	p.Asn863Valfs*4	LoF	yes	yes	yes
Δ(E7p25)	acceptor shift	r.2587_2611del	p.Asn863Metfs*1	LoF	yes	yes	yes
▼(E8p30)	acceptor shift	r.2748_2749ins2749-30_2749-1	p.Val917_Ser1186delins9	LoF	yes	-	yes
Δ(E8)	cassette	r.2749_2834del	p.Val917Glyfs*6	LoF	yes	yes	yes
Δ(E8_E9)	multi-cassette	r.2749_2996del	p.Val917Argfs*10	LoF	yes	yes	-
Δ(E10p2)	acceptor shift	r.2997_2998del	p.Gly1000Glnfs*9	LoF	yes	-	-
Δ(E10q31)	donor shift	r.3083_3113del	p.Thr1029Ilefs*1	LoF	yes	yes	-
▼(E11p23)	acceptor shift	r.113_3114ins3111-23_3114-1	p.Trp1038Cysfs*7	LoF	yes	yes	yes
Δ(E11p2)	acceptor shift	r.3114_3115del	p.Trp1038Ter	LoF	yes	yes	yes
Δ(E11)†	cassette	r.3114_3201del	p.Asn1039Glyfs*5	LoF	yes	yes	yes
Δ(E11)+▼(E12p446)	mixed	r.3114_3201del+r.3201_3202ins3202-446_3202-1	p.Trp1038Cysfs*3	LoF	yes	-	-
Δ(E11)+▼(E12p65)	mixed	r.3114_3201del+r.3201_3202ins3202-65_3202-1	p.Trp1038Ter	LoF	yes	-	-
▼(E12p65)	acceptor shift	r.3201_3202ins3202-65_3202-1	p.Gly1068Ilefs*28	LoF	yes	yes	yes
Δ(E12p136)	acceptor shift	r.3202_3337del	p.Leu1069Argfs*9	LoF	yes	-	-
Δ(E12)†,¶	cassette	r.3202_3350del	(p.Gly1068_Ser1186delins4)	LoF	yes	yes	yes

[1]See methods. [2]Biotype according to ENCODE [313]. [3]RNA described according to the Human Genome Variation Society rules at http://varnomen.hgvs.org/, using Ensembl transcript ENST00000261584.8 as a reference. ¶Described in comprehensive gene annotation from GENCODE release 26 retrieved through Ensembl at http://www.ensembl.org/. # The predicted 36 nucleotides insertion includes an in-frame PTC (p.Arg37_Ser1186delinsKTYFWGCFCLL). ## The predicted 30 nucleotides insertion includes an in-frame PTC (p.Val917_Ser1186delinsHNFWLLCFI). †described previously in the literature [131].

Targeted RNAseq data (Supplemental Table 1, Laboratory 1) indicated that most *high-confidence* events make on average (N=72 samples) a minor contribution to the expression level (reads supporting the splicing event representing ≤1% of the reads supporting the corresponding reference transcript). The only exceptions were Δ(E1q17), IVS1-463▼(134), Δ(E7p10), Δ(E11), Δ(E11_E12) and Δ(E12), with contributions of ≈2%, ≈5%, ≈1.4%, ≈2%, ≈2%, and ≈13% respectively). *In silico* analysis suggests that events contributing >1% might be related to the presence of sub-optimal splice sites at the *PALB2* gene (Supplemental Figure 7), with Δ(E12) contribution (≈13%) probably explained by the intrinsically weak exon 12 GC donor site [315]. The relatively elevated level of alternative splicing resulting in skipping of exons 11 and/or 12 is supported by targeted and whole transcriptome RNA-seq (Supplemental Table 1), semi-quantitative RT-PCR/CE analysis (Supplemental Figure 1A), whole-gene CloneSeq splicing analysis (Supplemental Figure 1B), and quantitative dPCR (Supplemental Figure 5B). According to the latter, ≈8%-34% of the *PALB2* transcripts (depending on the sample analyzed) may skip exon 11, exon 12, or both.

Overall coverage in whole transcriptome RNA-seq was substantially lower than in targeted RNA-seq experiments (Figure **53**). As a result, several events representing ≤1% of the targeted RNA-seq reads were not detected by this approach. Only one major discrepancy was observed related to *PALB2* Δ(E4_E5), which represented ≤1% of the corresponding reference signal in targeted RNA-seq and whole-exon GenClone experiments, but >5% in RNA-seq data generated by laboratory 3. However, subsequent digital PCR quantification in BLOOD, BREAST and OVARY confirmed that Δ(E4_E5) represents, on average, ≤1% of the corresponding reference signal (Supplemental Figure 5). Despite the lower coverage, whole transcriptome RNAseq and/or RT-PCR/CE experiments allowed us to detect 50 splicing events in BREAST, and 29 in OVARY. Of these, 24 splicing events-among them Δ(E1q17), IVS-463▼(134), Δ(E7p10), Δ(E11), Δ(E11_E12) and Δ(E12)-were detected in both tissues (Table 10 and Supplemental Table 1). Equally relevant, we did not identify tissue-specific *PALB2* alternate gene transcripts (neither in BREAST, nor in OVARY), suggesting that if they exist, they are expressed at very low levels -supporting the clinical relevance of BLOOD-based *PALB2* splicing studies.

Finally, we used data on alternate gene transcripts to analyze if PVS1 is warranted for all possible PTC-NMD/splice site variants at the *PALB2* gene. In brief, we concluded that PVS1 is warranted for every possible PTC-NMD variant, regardless of the location-i.e. we have not identified any plausible *rescue transcript* (see 'Discussion' section). By contrast, we conclude that PVS1 is not necessarily warranted for every possible splice site variant. To be more precise, we propose that PVS1 may not be warranted for splice site variants located at the acceptor sites of exons 2, 5, 7 and 10. For this subset of splice site variants, the production of RNA transcripts retaining some or all functional capacity is plausible (see Table 12 for further details). If splicing assays and/or clinical data supporting pathogenicity is lacking, we recommend caution when classifying splice site variants at these specific sites-i.e. such variants should not be assumed *pathogenic/likely pathogenic*.

Table 12 : Proposed classification of PALB2 splice site variants according to the ACMG-AMP-2015 guidelines (based solely on location and MAF)

Splice Site	Variant	Predicted RNA Products /Coding Potential[1]		PVS1[1]	gnomAD[2]	PM2[2]	Classification[3]
		LoF [1]	**Uncertain[1]**				
E1 donor	c.48+1,2	Δ(E1q17)†	-	warranted	-	yes	Likely pathogenic
E2 acceptor	c.49-1,2	-	Δ(E2p6)†	not warranted	NFE (1 allele)	yes	**Uncertain Significance**
E2 donor	c.108+1,2	Δ(E2)/▼(I2)	-	warranted	-	yes	Likely pathogenic
E3 acceptor	c.109-1,2	▼(E3p36)/Δ(E3)	-	warranted	-	yes	Likely pathogenic
E3 donor	c.211+1,2	Δ(E3)	-	warranted	-	yes	Likely pathogenic
E4 acceptor	c.212-1,2	Δ(E4_E5)†	-	warranted	NFE (1 allele)	yes	Likely pathogenic
E4 donor	c.1684+1,2	Δ(E4_E5)†	-	warranted	-	yes	Likely pathogenic
E5 acceptor	c.1685-1,2	Δ(E5)	Δ(E5p24)	not warranted	NFE (1 allele)	yes	**Uncertain Significance**
E5 donor	c.2514+1,2	Δ(E5)	-	warranted	SAS (1 allele)	yes	Likely pathogenic
E6 acceptor	c.2515-1,2	Δ(E6)†	-	warranted	AMR (1 allele)	yes	Likely pathogenic
E6 donor	c.2586+1,2	Δ(E6)†	-	warranted	SAS (1 allele)	yes	Likely pathogenic
E7 acceptor	c.2587-1,2	▼(E7p20)/Δ(E7p2)/Δ(E7p10)/Δ(E7p25)/Δ(E7)	▼(E7p42)	not warranted	SAS (1 allele)	yes	**Uncertain Significance**
E7 donor	c.2748+1,2	Δ(E7)†	-	warranted	NFE (1 allele)	yes	Likely pathogenic
E8 acceptor	c.2749-1,2	▼(E8p30)/Δ(E8)	-	warranted	-	yes	Likely pathogenic
E8 donor	c.2834+1,2	Δ(E8)	-	warranted	-	yes	Likely pathogenic
E9 acceptor	c.2835-1,2	Δ(E9p30)†/Δ(E9)†	-	warranted	-	yes	Likely Pathogenic
E9 donor	c.2996+1,2	Δ(E9)/Δ(E9_E10)	-	warranted	-	yes	Likely pathogenic
E10 acceptor	c.2997-1,2	Δ(E10p2)/Δ(E9_E10)/Δ(E10)	Δ(E10p3)	not warranted	SAS (1 allele)	yes	**Uncertain Significance**
E10 donor	c.3113+1,2	Δ(E10q31)†/Δ(E9_E10)†/Δ(E10)†	-	warranted	-	yes	Likely pathogenic
E11 acceptor	c.3114-1,2	Δ(E11)/Δ(E11p2)/Δ(E11p23)/Δ(E11_E12)	-	warranted	-	yes	Likely pathogenic
E11 donor	c.3201+1,2	Δ(E11)/Δ(E11_E12)	-	warranted	-	yes	Likely pathogenic
E12 acceptor	c.3202-1,2	▼(E12p65)/Δ(E12p136)/Δ(E11_E12)/Δ(E12)	-	warranted	-	yes	Likely pathogenic
E12 donor	c.3350+1,2	Δ(E11_E12)†/Δ(E12)†	-	warranted	-	yes	Likely pathogenic
E13 acceptor	c.3351-1,2	-	-	warranted	-	yes	Likely pathogenic

[1] If available (†), predictions on possible RNA products are based on splicing assays performed in representative examples of splice site variants (see Supplemental Table 4). If not, predictions are based on the possible up-regulation of naturally occurring alternate gene transcripts. Predicted RNA products are classified according to their coding potential as loss-of-function (LoF) or uncertain (the possibility of coding for a functional or partially functional protein cannot be disregarded). If only LoF transcripts are predicted, we assume that PVS1 is warranted. If ≥1 transcript with uncertain coding potential is predicted, we propose that PVS1 (based solely on variant location) is not warranted. [2]After reviewing gnomAD, we conclude that PM2 is met for all possible splice site variants. [3]According to the ACMG-AMP-2015 guidelines, if PVS1 and PM2 are warranted, splice site variants should be classified as likely pathogenic. Otherwise, splice site variants should be classified as uncertain significance. This analysis has highlighted 7 splice site variants in ClinVar needing additional justification for assertion as Pathogenic/Likely Pathogenic (see Supplemental Table 5 for further details). NFE (non-finish Europeans). SAS (South Asia). AMR (American)

5. Discussion

Alternative splicing probably occurs in all metazoan organisms, and increasing prevalence has been linked to phenotypic complexity [316]. Virtually all human multi-exon *loci* produce alternate gene transcripts [317]. Apart from a presumed role in expanding protein diversity [318] that is currently under dispute [319], [320], some authors have suggested that alternative splicing may buffer mutational consequences [321]. The latter possibility has obvious implications for the clinical interpretation of genetic testing results. The ACMG-AMP-2015 guidelines acknowledge this by recommending caution about over-interpreting the impact of PTC-NMD and splice site variants if multiple transcripts are present [232]. Here we have addressed this relevant aspect of alternative splicing for the particular case of classifying genetic variants at the breast cancer predisposition gene *PALB2*.

Alternative splicing analysis might be influenced by many factors, including collection of RNA samples, experimental design, and detection sensitivity. For instance, one study characterizing alternative splicing at breast cancer susceptibility genes by RNA-seq noticed the poor performance of PAXgene if compared with LCL samples [131], and a previous ENIGMA collaborative study comparing RT-PCR splicing protocols across different laboratories concluded that primers design and detection sensitivity (rather than RNA extraction and/or cDNA synthesis protocols) had an impact on the analytical outcome [322]. A strength of our study design was the application of different assay designs, RNA samples and subsequent levels of sensitivity and/or filtering, by five independent laboratories to identify PALB2 alternative splicing events (see online supplementary material section 1 for further details). We elected to define high-confidence splicing events as those found in at least two different data sets (the rationale being that events detected by a minimum of two laboratories, two sample types and two methodologies are very unlikely to represent technical artefacts and/or biological outliers), but acknowledge that such definition may lead to exclusion of real events found by a single laboratory. A higher stringency of high-confidence splicing events found by more than two laboratories was not used due to differences in the level of sensitivity between assays.

Overall, we identified 44 high-confidence alternative splicing events at the PALB2 locus, and we provide evidence for 44 additional events (although we cannot discard the possibility that some of the latter represent technical artefacts and/or biological outliers). Interestingly, all PALB2 reference exons are affected by one or more high-confidence alternative splicing events, suggesting that no PALB2 exon should be annotated as constitutive. Despite the considerable number of alternative splicing events identified, our data suggest that their contribution to the overall PALB2 expression is low in all three tissues investigated. Splice site and PTC-NMD variants in cancer susceptibility genes can be overinterpreted (misinterpreted as pathogenic), if alternate gene transcripts are not properly considered. [232], [255], [312], [323]–[325]. In the past, this has led to errors in the clinical management of families carrying the *BRCA1* allele c.[594-2A>C; 641A>G] [325]. The low level of alternative splicing observed

for *PALB2* in BLOOD, BREAST, and OVARY suggests that overinterpreting genetic variants at this locus is less likely to occur. However, some of the alternative splicing events we report can be relevant for the clinical interpretation of *PALB2* PTC-NMD and splice site variants, in particular to decide if PVS1 is warranted.

PTC-NMD variants: the existence of *rescue* transcripts reducing or eliminating the functional and clinical impact of certain PTC-NMD variants in cancer susceptibility genes has been confirmed for *APC* [324] and *BRCA1* [255]. More specifically, the alternate gene transcript APC Δ(E9p303) explains the association of PTC-NMD variants located at codons 312-412 with mild disease [324], and the alternate gene transcript *BRCA1* Δ(E9_E10) explains the low breast cancer risk observed in carriers of the splice site variant *BRCA1* c.594-2A>C [255]. However, we have not identified plausible rescue transcripts for *PALB2*. Alternate gene transcripts Δ(E2p6), Δ(E6), Δ(E5p24) and Δ(E10p3) might code for functional or partially functional proteins, but their respective contribution to the overall *PALB2* expression (<1%) is too low to be plausible *rescue* transcripts. By contrast, the combined expression of Δ(E11_E12) and Δ(E12) might represent 8-34% of the overall gene expression (depending on samples and methodologies), but the predicted proteins encoded by these two transcripts (Table 10) are unlike to be functional, as they lack part of the C-terminal WD40 β-propeller domain (see Supplemental Material section 2.3) that mediates PALB2 interaction with several key homologous recombination proteins, including BRCA2 and RAD51[326]. For that reason, we do not consider Δ(E11_E12) and Δ(E12) plausible rescue transcripts, although we cannot rule out the possibility of truncating variants in exons 11 and/or 12 conferring lower cancer risk than truncating variants in other *PALB2* exons.

Canonical ±1,2 splice site variants: we propose that naturally occurring alternate gene transcripts provide predictive information identifying seven *PALB2* canonical splice sites for which, in absence of splicing assays, PVS1 is not warranted (variants targeting exons 2, 5, 7, and 10 acceptor sites). For exon 2 acceptor site, the proposal is based on experimental data obtained in a *PALB2* c.49-1G>A (IVS1-1G>A) carrier indicating up-regulation of Δ(E2p6) (Dr. Georgios Tsaousis, Genekor Medical S.A., personal communication, June2018). The possibility that Δ(E2p6) code for a functional/partially functional protein cannot be discarded (**see Supplemental Material section 2.3**), supporting our conservative stance. For the remaining splice sites, we hypothesize that naturally occurring alternate gene transcripts (even if lowly expressed in control samples) may become up-regulated if splice site variants impair the expression of reference transcripts. The hypothesis is supported by several observations made in carriers of *PALB2* (among them, the up-regulation of Δ(E2p6) in c.49-1G>A carriers), *BRCA1*, and *BRCA2* splice site variants **(see Supplemental Table 4).** Note that we propose that PVS1 is not warranted for splice site variants if at least one RNA product with uncertain coding potential is predicted, regardless of other predictions. For instance, we propose that PVS1 is not warranted for variants targeting the *PALB2* exon 7 acceptor site because one RNA product of uncertain coding potential, ▼(E7p42), is predicted (Table 12), despite the fact that up to five bona fide LoF

transcripts are also predicted (▼(E7p20), Δ(E7p2), Δ(E7p10), Δ(E7p25),and Δ(E7)). When classifying splice site variants in high risk breast cancer genes as pathogenic/likely pathogenic without functional or genetic data, we favor a very conservative approach. We have identified 43 different *PALB2* splice site variants in ClinVar (last accessed 13/04/2018), all them reported as pathogenic/likely pathogenic. For 4 of these variants, we think that the pathogenic/likely pathogenic classification may not be justified without considering additional clinical and/or splicing data (Table 13).

In short, we highlight the fact that, where alternate gene transcripts exist, assertions of pathogenicity are warranted only with the support of additional quantitative splicing assays, and preferably clinical evidence.

Table 13 : Known PALB2 splice site variants for which we put a warning.

			ClinVar			
Splicing Site	**Variant reported**	**dbSNP**	**Classification**	**Review status**	**Assertion method**	**Proposed ACMG-2015 Classification**
E2 acceptor	c.49-2A>T	rs786203245	Likely Pathogenic	**	Ambry autosomal dominant Invitae Variant Classification Sherlock	Uncertain Significance
E5 acceptor	c.1685-2A>G	rs754660432	Likely Pathogenic	**	GeneDx variant Classification Ambry autosomal dominant	
	c.1685-1G>C	rs1057520645	Pathogenic	*	GeneDx variant Classification	
E7 acceptor	c.2587-2A>C	rs1060502787	Likely Pathogenic	*	Invitae Variant Classification Sherlock	
E10 acceptor	c.2997-2A>C	-	Likely Pathogenic	*	Ambry autosomal dominant	

These five *PALB2* variants are classified as pathogenic/likely pathogenic based on assertion criteria defined by the submitters. Ambry Genetics and/or GeneDx classify the indicated variants as pathogenic based on the fact that these are very rare variants located at canonical splice sites, predicted to abolish or significantly reduce native site using in silico predictors and identified in affected/+family history cohort. Invitae classifies the indicated variants as likely pathogenic based on the fact that donor and acceptor splice site variants are typically loss-of-function and loss-of-function variants in *PALB2* are known to be pathogenic. Remarkably, for any of these variants classification is based on splicing assays, and/or in segregation information supporting pathogenicity (Tina Pesaran, unpublished data; Kathleen S Hruska, unpublished data, Inviate ClinVar summary evidences). These are splice site variants targeting acceptor sites for which, in our opinion (Table 12), PVS1 is not necessarily warranted. For that reason, we propose that, in absence of functional and/or genetic data, these variants should be classified according to the ACMG-AMP-2015 guidelines as uncertain significance. ACMG-AMP, American College of Medical Genetics and Genomics-Association for Molecular Pathology.

6. Declaration

a. Acknowledgments

The authors would like to thank A Leconte at cancer center, F Baclesse and to Dr C Baudouin at Polyclinique du Parc, Caen (France) for their participation to CASOHAR clinical trial to obtain breast tissue from healthy volunteers. The authors would like to thank Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics and the Clinical Follow-Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia and the National Institute of Health [USA]) for their contributions to this resource, and the many families who contribute to kConFab. The authors would like to thank the SCAN-B collaborators at participating hospitals for support of SCAN-B. The authors would also like to thank Ingrid Hedenfalk and staff at the collaborating gynaecology and pathology clinics in Lund for providing the fimbriae samples.

b. Contributors

IL-P, RL, RB, VL, JFP, LC, AM, DV, NG, GD, PG, VG-B, PLl, PP-S, ED-R, TC, KSH, VH, SW, TP, RK, JV-C, AV-P and MS, contributed to data acquisition, revised the manuscript for important intellectual content and approved the final version. kConFAB provided research resources used in this study. ABS coordinated the ENIGMA consortium. AB, EAV, MPV, PD, AK, ABS, LW and SK contributed to the conception and design of the study, contributed to obtain all necessary approvals and clearances to conduct the research, contributed to data acquisition, contributed to data analysis, contributed to grant funding, revised the manuscript for important intellectual content and approved the final version. MdH contributed to obtain all necessary approvals and clearances to conduct the research, contributed to the conception and design of the study, contributed to data acquisition, contributed to data analysis, contributed to grant funding, wrote the manuscript and approved the final version.

c. Funding

PD, MPGW, EAV, AB, AK and MH have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 634935. RL is supported by a Normandy-University, Federation-Hospitalo-Universitaire (FHU) grant. VL is supported by a Mackenzie Family Cancer Postdoctoral Fellowship. RB is supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 634935. AM is supported by a French Cancéropôle Nord-Ouest (CNO) grant. AB is supported by Mrs Berta Kamprad Foundation. EAV and MH are supported by Spanish Instituto de Salud Carlos III (ISCIII) funding (grants PI17/00227 to EAV and PI15/00059 to MH), an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds. ABS is supported by an NHMRC Senior Research Fellowship (ID1061779). LCW is supported by the Rutherford Discovery Fellowship. SK is

supported by Ligue Contre le Cancer, Normandie. kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia and the Cancer Foundation of Western Australia.

d. Competing Interests.

VH, SW, PT, RK were employees of Ambry Genetics when they were engaged with this project. KSH was employee of GeneDx when she was engaged with this project. EDR has consulting or advisory roles in Amgen, Bayer, Genómica, Servier and Merck. EDR has got research funding from: Roche, Merck-Serono, Amgen, AstraZeneca and Sysmex.

e. Ethics approval

Ethics approval Academic Hospital San Carlos ethics committee (reference numbers 15/139 E and 16/505 E). The SCAN-B study has been approved by the Lund Regional Ethical Review Board, Sweden (approval number 2009/658). The fimbriae tissue samples were obtained and analysed with approval by the Lund Regional Ethical Review Board, Sweden (approval number 2014/717). French Biomedicine Agency. CASOHAR trial ethic committee (NTC NTC02560818). Ambry Genetics' patient's information has been de-identified, and this study has been approved and carried out in accordance with the recommendations of the Western Institutional Review Board (WIRB; IRB Tracking Number:20171324). Whole-transcriptome RNAseq study was approved by the New Zealand Southern Health and Disability Ethics Committee (12/STH/44).

f. Data sharing

Targeted RNAseq data contributed by laboratory 1 is available from Dr Sophie Krieger on reasonable request. Whole-transcriptome RNAseq data generated by laboratory 3 is available from Dr Logan Walker on reasonable request. Targeted RNAseq data contributed by laboratory 2 is available from SCAN-B and Ingrid Hedenfalk, respectively, but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are however available from the authors on reasonable request and with permission of SCAN-B or Ingrid Hedenfalk.

DISCUSSION

Au cours de cette thèse nous avons abordé trois aspects critiques de l'étude de l'impact des variants sur l'épissage : la prédiction des défauts d'épissage, l'analyse des données de RNA-seq et l'importance de l'épissage dans l'interprétation à usage clinique des variants.

I. Les prédictions des défauts d'épissage : les avancées et limites

Le nombre croissant de variants nucléotidiques détectés par NGS a eu pour conséquence l'essor d'outils *in silico* de prédiction pour aider à interpréter ces variants. La prédiction des défauts d'épissage est un parfait exemple de cet essor au vue du nombre d'outils publiés au cours de ces dernières années.

1. Quels outils de prédiction pour quels motifs d'épissage

Ainsi, lors de ce travail de thèse, nous avons dans un premier temps réévalué les recommandations existantes concernant l'utilisation de ces prédictions d'une altération des sites consensus donneur/accepteur publiées par le GGC [254]. Ce travail a conduit au développement d'un nouvel outil : SPiCE.

SPiCE combine les deux outils de prédiction MES et SSF. SPiCE a démontré des performances supérieures à l'utilisation des scores MES et SSF seuls ainsi qu'à ces précédentes recommandations. Par ailleurs ses performances sont maintenues lorsque SPiCE est évalué sur d'autres variants que *BRCA1/BRCA2*. A partir des scores fournis par SPiCE, nous avons dégagé deux seuils décisionnels pour prioriser ou non une étude ARN *in vitro*. Le premier seuil a été conçu pour maitriser le risque de faux négatifs en définissant une sensibilité optimale. En effet dans un contexte de diagnostic moléculaire, il est crucial de ne pas exclure à tort des variants splicéogéniques, ces derniers pouvant être pathogènes. Le second seuil décisionnel a été défini par rapport à une spécificité optimale dans le but de prioriser au titre de la recherche ces variants. Ces deux seuils peuvent aussi s'utiliser de façon concomitante avec la création d'une zone grise entre eux. Néanmoins nous avons observé que seulement 10 % des variants utilisés pour la validation de SPiCE étaient observés dans cette zone (16/160 et 9/90). Ainsi SPiCE s'est révélé être un outil particulièrement efficace pour prioriser les études ARN *in vitro* pour les variants situés dans les régions consensus d'épissage.

Cependant les motifs consensus d'épissage sont loin d'être les seuls signaux utilisés par le splicéosome. Les variants situés dans les motifs consensus donneur/accepteur d'épissage sont largement étudiés. Cependant pour les autres motifs d'épissage, l'impact de variants est bien moins décrit. Les points de branchement illustrent ce défaut. Actuellement, sur le site de Pubmed, 1 136 articles sont recensés avec les mots clés « *consensus site splice* » contre 340 avec les mots clés « *branch point splice* » (septembre 2019).

Aussi durant ce travail de thèse nous nous sommes dans un second temps focalisés sur l'étude des prédictions des points de branchement en comparant 6 outils : HSF, SVM-BPfinder, BPP, Branchpointer, LaBranchoR et RNABPS. Deux problématiques ont été considérées :

- La capacité des outils à identifier des points de branchement uniquement devant les sites accepteurs reconnus par le splicéosome.
- La capacité des outils à détecter une altération d'un point de branchement.

Ainsi nous avons pu identifier l'outil Branchpointer comme optimal pour discriminer la présence ou non d'un point de branchement. Cet outil met en valeur l'intérêt du *deep learning* sur lequel il repose, pour la détection des motifs d'épissage par rapport aux précédents algorithmes. En effet BPP, SVM-BPfinder et HSF, publiés antérieurement à Branchpointer, utilisent respectivement : un modèle mixte, SVM et PWM [146], [153], [261]. Néanmoins l'utilisation du deep learning ne permet pas à elle seule de justifier des performances de Branchpointer, car RNABPS et LaBranchoR utilisent également un tel algorithme [167], [168]. En réalité, Branchpointer intègre aussi la structure des transcrits dans le calcul du score tandis que RNABPS et LaBranchoR ne considèrent que la séquence nucléotidique. De plus Branchpointer a été entrainé uniquement sur les points de branchement associés aux transcrits majoritairement exprimés, nommés *high-confidence* par les auteurs [166]. Ce dernier point permet notamment d'expliquer l'excellente spécificité sur les données Ensembl (99.49 %). A l'inverse, Branchpointer ne parvient pas à détecter de points de branchement pour les transcrits faiblement exprimés observés par RNA-seq, sensibilité de 32.1 %. Ceci met en lumière l'importance du choix des stratégies d'entrainement des outils, notamment pour ceux basés sur le *deep learning*.

Nous avons également établi une des plus importantes collections de variants, situés dans la région des points de branchement, avec leurs études ARN *in vitro* (120 variants), grâce à une collaboration nationale au sein du GGC, de l'ANPGM et internationale *via* ENIGMA. Nous avons ainsi pu identifier que pour détecter l'altération d'un point de branchement l'outil Branchpointer n'est plus l'outil optimal. En effet, c'est l'outil BPP dans cette situation qui présente les meilleures performances. Par ailleurs, les variants splicéogéniques sont concentrés sur les motifs de points de branchement prédits par BPP, notamment sur le A du point de branchement et le T deux nt en amont. Il en résulte que nous pouvons prédire un variant comme splicéogénique s'il se situe dans le motif TRAY du point de branchement prédit [262].

En parallèle de l'étude des points de branchement, nos collègues Rouannais, de l'équipe Inserm U1245, ont grandement contribué à la caractérisation des ESRs. Ils ont ainsi étudié la relation entre la prédiction de ces motifs et les défauts d'épissage. Dans un premier temps les variants des gènes *MLH1*, *BRCA1*, *BRCA2*, *CFTR* et *NF1* ont été pris comme modèle d'étude [276]. Les outils de prédiction considérés étaient ΔtESRseq, ΔHZ_{EI} et SPANR [150], [173], [327]. Actuellement cette étude est étendue sur une

plus large cohorte de gènes [328]. Ces travaux ont ainsi montré que l'outil de prédiction des ESR, ΔtESRseq, est un des outils optimaux pour prédire un défaut d'épissage.

Ainsi la définition des outils *in silico* de prédiction optimaux offre une aide significative dans la démarche visant à l'interprétation des variants. Cependant la plupart de ces outils sont spécifiques d'un motif d'épissage ou d'un même groupe de motifs. Aussi se pose la question d'une méthode apte à prédire un défaut d'épissage indépendamment de la position du variant et du motif d'épissage pouvant être impacté.

Dans le but de répondre à cette question nous avons développé SPiP. En effet, cet outil s'adresse à la diversité des motifs d'épissage quel que soit la localisation du variant dans le gène. Grâce à la combinaison d'outils optimaux pour chaque motif, SPiP a atteint une exactitude de 80.21 % avec une sensibilité de 90.96 %, sur un jeu de 2 784 variants. En outre, cette sensibilité s'est révélée supérieure à celle obtenues par les deux récents outils de *deep learning*, SPIDEX (78.37 %) et SpliceAI (70.71 %). Par ailleurs, SPiP a été conçu pour rapporter quel est le motif altéré et la probabilité d'altération de l'épissage, afin de faciliter l'interprétation des résultats. Effectivement, la probabilité d'observer un défaut d'épissage varie en fonction du motif altéré. Ainsi nous avons observé une amplitude de 0.9 % à 90 %, respectivement pour la création d'un site d'épissage dans l'intron profond et l'altération d'un motif consensus du site naturel d'épissage. Au total, les motifs dont l'altération induit une probabilité élevée de défaut d'épissage, sont ceux contribuant à définir les jonctions exon/intron : motif consensus donneur/accepteur, tract polypyrimidique, point de branchement et ESRs. A l'inverse la création d'un nouveau motif d'épissage, notamment à distance des sites naturels, présente un plus faible risque de modifier l'épissage. Le calcul de ces probabilités, nous a aussi révélé que les prédictions négatives de SPiP permettent d'exclure les variants pour une étude ARN *in vitro*. Par conséquent, SPiP a le potentiel d'être un outil de décision à large échelle pour aiguiller le généticien vers les variants splicéogéniques.

2. Faut-il se limiter à la seule prédiction d'une altération de l'épissage

Il existe une grande variété d'outils de prédiction pour l'épissage, et au cours de ce travail de thèse nous avons pu proposer un ensemble d'outils qualifiés comme optimaux pour prédire un défaut d'épissage. Cependant un point majeur à souligner est que ces outils prédisent l'altération/création des motifs d'épissage et non la nature du défaut d'épissage. Cette différence entre ces deux notions s'illustre particulièrement pour les sites consensus donneur/accepteur. En effet, notre outil SPiCE a montré une exactitude de 95.6 % pour prédire un défaut d'épissage. Néanmoins, même au sein des vrais positifs de SPiCE du jeu de validation (n = 207), nous pouvons observer aussi bien un saut d'exon (156/207) que l'utilisation d'un nouveau site d'épissage (51/207). De plus parmi les 2 784 variants utilisés pour développer SPiP, nous avons aussi pu constater que des variants situés dans les sites consensus étaient

à l'origine d'une rétention complète d'intron (n = 4) et l'apparition de pseudo-exon (n = 2). Une partie de cette diversité peut s'expliquer par la présence de l'épissage alternatif comme nous avons pu le montrer pour le gène *PALB2* [264]. Cependant, les variants de *BRCA2* : c.8754G>A ; c.8754G>C ; c.8754+1G>A ; c.8754+1G>C ; c.8754+3G>C ; c.8754+4A>G ; c.8754+5G>A ; c.8754+5G>T, tous situés dans le site consensus donneurs de l'exon 21, entrainent systématiquement l'utilisation d'un site donneur à 46 nt dans l'intron [254], [260], [303], [304]. Or cette rétention des 46 premières nt ne correspond à aucun épissage alternatif connu de *BRCA2* [131], [191].

Aussi certains outils de prédiction se proposent de rechercher la présence d'un site alternatif pouvant être utilisé en relai du site naturel. Par exemple, l'outil CRYP-SKIP calcule la probabilité d'utilisation d'un site cryptique au lieu du saut d'exon [171]. Cependant CRYP-SKIP a une utilisation restreinte puisqu'il n'est applicable que pour les variants altérant l'épissage. Plus récemment, nous pouvons également citer SpliceAI pour répondre à cette question. En effet, cet outil fournit deux scores, un pour la perte des sites d'épissage et le second pour le gain d'un autre site d'épissage [298]. Aussi, si nous étudions les scores de SpliceAI pour les variants consensus et splicéogéniques (n = 783) issus des 2 784 variants de SPiP. Parmi ces 783 variants, 288 variants entrainent l'utilisation d'un nouveau site d'épissage. SpliceAI détecte ces évènements avec une exactitude de 75.9 % (595/783), une sensibilité de 55.2 % (159/288) et une spécificité de 88.1 % (436/495). L'utilisation d'un tel outil récent basé sur le *deep learning* ne permet de détecter guère plus de la moitié de l'utilisation d'un nouveau site d'épissage lorsque le site naturel est altéré. De plus, ni CRYP-SKIP, ni SpliceAI n'ont été conçus pour détecter l'utilisation d'un pseudo-exon ou la rétention complète d'intron lorsque le site naturel est endommagé.

Nous avons également pu remarquer cette diversité des défauts d'épissage pour l'altération d'un même type de motif dans le cadre de l'étude des prédictions des points de branchement. En effet parmi les 120 variants collectés pour cette étude, 38 variants modifiaient l'épissage dont 7 entrainaient l'utilisation d'un site accepteur alternatif à distance du site naturel. A titre d'exemple, le variant *KCNH2* c.2399-28A>G induit l'utilisation d'un site accepteur alternatif à 147 nt en amont du site naturel [329]. Par ailleurs grâce aux 1 294 variants splicéogéniques rapportés dans la collection des 2 784 variants de SPiP, nous avons observé que les rétentions complètes d'intron s'observaient fréquemment lorsque le point de branchement était altéré. Ainsi, sur les 16 variants induisant une rétention complète d'intron, 7 sont prédits comme altérant un point de branchement. Cependant pour prouver l'association entre ces deux évènements, des études complémentaires sont nécessaires.

Outre la diversité des altérations, un allèle muté peut produire différents transcrits en proportion variable. En effet, un variant peut engendrer à la fois un transcrit aberrant et le transcrit naturel, ce phénomène est alors qualifié d'effet partiel. Ainsi se pose la question de savoir si un score de prédiction d'altération de l'épissage est corrélé au niveau d'expression de ces différents transcrits. Lors du développement de

SPiCE, nous avons pu constater que la majorité des variants ayant un effet partiel a un score intermédiaire, c'est-à-dire la classe medium de SPiCE. Durant l'étude des prédictions des points de branchement, les sites accepteurs alternatifs, identifiés par RNA-seq, ont un niveau d'expression significativement plus élevé lorsqu'un point de branchement est prédit en amont du site. Cependant cette corrélation reste plus que faible au regard du coefficient de détermination (R^2) maximal de 0.0062 pour RNABPS. Effectivement, l'expression d'un transcrit dépend d'un certain nombre de facteurs. Parmi lesquels nous pouvons mentionner de manière non exhaustive :

- La présence d'autres motifs d'épissage, les séquences des sites donneurs/ accepteurs ou bien la composition en séquences régulatrices ESRs.
- L'existence d'autres variants pouvant modifier l'impact d'un variant sur l'épissage [255].
- Différents facteurs régulant la transcription et la traduction, par exemple les snRNAs et les micro ARNs.
- L'expression tissu-spécifique des gènes comme nous l'a montré le projet GTEx (*Genotype-Tissue Expression*) (https://gtexportal.org/home/) [330].
- Le recrutement du NMD pour les transcrits avec un PTC.
- La présence des épissages alternatifs.

De surcroit d'autres études de développement d'outils se sont proposés de corréler leurs scores au niveau d'expression des transcrits. Parmi ces derniers, nous pouvons citer un outil récemment publié MMSplice (*modular modeling of splicing*) [293]. Le score MMSplice est calculé par un algorithme de *neural network* à partir des séquences exoniques et introniques bordant le variant. Les auteurs de MMSplice ont confronté leur score aux données d'expression du projet Vex-seq [100]. Le projet Vex-seq consistait à étudier par RNA-seq le niveau d'inclusion d'un exon dans les transcrits issus de tests minigènes à haut débit. Ainsi les auteurs du projet Vex-seq ont pu caractériser l'impact de 2 059 variants sur l'inclusion de 110 exons. En outre, la construction artificielle des minigènes a permis de réduire l'impact d'éventuel polymorphisme, de facteur de transcription, de l'expression tissu-spécifique, du NMD et des épissages alternatifs. Aussi ces données présentent des conditions optimales pour tester la corrélation entre un score de prédiction et le niveau d'expression des transcrits. Bien que MMSplice ait montré une corrélation plus importante que ses prédécesseurs comme SPANR ou HAL [25], [173]. Il n'a pas atteint un coefficient de détermination supérieur à 0.5 ($R^2 = 0.46$). En d'autre terme, même dans ces conditions optimales et avec un outil surpassant ses prédécesseurs, plus de la moitié de la variabilité d'expression d'un transcrit n'est pas expliquée par les scores de prédiction.

A l'ère du séquençage NGS, les prédictions de l'altération des motifs d'épissage apportent une aide capitale pour détecter et prioriser les variants splicéogéniques. Toutefois, la somme de ces arguments indique que ces prédictions ne peuvent se substituer aux études ARN *in vitro* pour préciser la nature et le degré d'altération du défaut d'épissage, afin de conclure sur la pathogénicité le cas échéant.

II. L'apport du RNA-seq dans l'étude des modifications d'épissage

La nécessité de caractériser expérimentalement les défauts d'épissage pour un nombre grandissant de variants splicéogéniques a constitué un terrain favorable au développement du RNA-seq. Cependant si plusieurs outils sont disponibles pour détecter une différence d'expression de gènes ou d'exon, peu d'outils sont disponibles pour identifier et quantifier les jonctions d'épissage.

1. Identification des évènements d'épissage à partir de données RNA-seq

La caractérisation des épissages alternatifs et des défauts d'épissage peut s'avérer cruciale pour l'établissement d'un diagnostic moléculaire. Aussi au cours de ce travail de thèse nous avons développé SpliceLauncher pour étudier ces épissages [331]. SpliceLauncher s'est avéré capable de caractériser plus de 90 % des épissages alternatifs décrits dans le gène *PALB2* [264]. De plus, SpliceLauncher a permis de confirmer des défauts d'épissage difficilement identifiables par RT-PCR, comme le saut de l'exon 11 et la Δ11q(3309) causés par le variant c.4096+3A>T du gène *BRCA1*. SpliceLauncher comprend un pipeline bioinformatique pour extraire les données des jonctions d'épissage à partir des fichiers bruts FastQ. De plus, au travers de cet outil, nous proposons le moyen de visualiser ces jonctions d'épissage d'une part par l'intermédiaire de l'UCSC *Genome Browser, via* la génération de fichiers BED, et d'autre part par la génération de fichiers PDF permettant par gène de visualiser dans son ensemble ces jonctions. Pour faciliter la lecture et l'interprétation des résultats, SpliceLauncher détermine logiquement la nature de la jonction d'épissage et ses coordonnées transcriptomiques. En outre l'analyse statistique réalisée par SpliceLauncher est suffisamment sensible pour détecter l'expression significativement différente supportée par un seul échantillon. La seule contrainte pour cette analyse est d'avoir suffisamment de données par jonction pour modéliser une loi de probabilité (gamma ou binomiale négative). Cependant SpliceLauncher nécessite un nombre restreint d'échantillon aux regards des autres outils d'analyse. En effet la plupart des outils statistiques en RNA-seq (ex : DESeq2) nécessite le regroupement des échantillons en différents ensembles [332]. Usuellement il est recommandé de disposer d'au moins cinq réplicas par groupe [140]. Ainsi, à titre d'exemple, pour analyser l'épissage de 9 patients porteurs d'un variant plus un contrôle, il faudrait pour une approche de type DESeq2 avoir au moins cinq échantillons biologiques pour chaque patient. Puis il serait nécessaire de séquencer 50 échantillons $(5 \times (9 + 1))$ simultanément. Alors que SpliceLauncher ne nécessite qu'un réplica par patient, soit une réduction de 50 à 10 échantillons à séquencer.

L'analyse de SpliceLauncher offre la possibilité de quantifier et de détecter une expression anormale de jonction d'épissage, comme a pu le montrer l'analyse du variant du gène *BRCA1* c.4096+3A>T (Figure 54). Néanmoins, nous ne disposons pas de seuil décisionnel pour considérer un effet comme partiel ou total. En effet à ce jour, seuls les travaux de l'équipe de Rien Blok ont pu montrer une distinction entre

un effet partiel et total grâce à un outil développé par cette équipe : QURNAS (non publié) et aux données de RNA-seq ciblé des gènes *BRCA1*/*BRCA2* et *RAD51C*/*D* [333]. Cependant l'identification et la quantification des évènements d'épissage semblent étroitement liées au choix du protocole expérimental, de la plateforme de séquençage et du pipeline bioinformatique utilisés. De ce fait l'identification d'un évènement partiel ou total est conditionnée par la méthodologie employée. Nous disposons actuellement de comparaison et de recommandation pour l'évaluation des défauts d'épissage pour les technologies d'analyses à bas débit (ex : RT-PCR et test minigène) [97], [254]. Cependant de telles recommandations n'existent pas pour le RNA-seq.

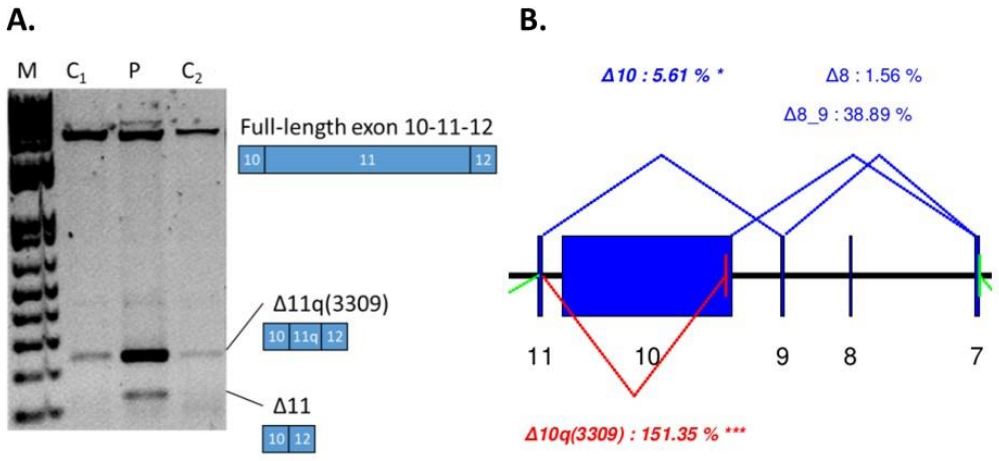


Figure 54 : Comparaison entre RT-PCR et RNA-seq pour la détection et quantification des évènements d'épissage du variant c.4096+3A>T dans le gène *BRCA1*. **A.** RT-PCR long-range, M : échelle, C_1/C_2 : Témoins non mutés, P : patient porteur du variant c.4096+3A>T dans le gène *BRCA1*. **B.** Données RNA-seq analysées par SpliceLauncher, telles que représentées dans les fichiers PDF générés par cet outil. Les numéros d'exon sont décalés en raison de l'absence d'exon 4 pour le gène *BRCA1*

2. Comparaison des analyses RNA-seq

Face au besoin d'adapter les technologies de RNA-seq pour l'étude des défauts d'épissage et de l'absence de recommandation pour interpréter ces données, le consortium ENIGMA se propose d'évaluer un panel de technologies RNA-seq. Nous participons au projet *quality control RNA-seq* ou QC RNA-seq, piloté par Logan Walker (*Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand*). Ce projet a pour objectif de comparer les performances de ces différentes technologies dans l'évaluation des défauts d'épissage. Pour mener à bien ce projet les lignées lymphoblastoïdes de patients porteurs des variants : *BRCA1* c.135-1G>T, c.591C>T, c.594-2A>C, c.671-2A>G, c.5467+5G>C et *BRCA2* c.426-12_8delGTTTT, c.7988A>T, c.8632+1G>A, c.9501+3A>T ainsi que 10 lignées lymphoblastoïdes contrôles ont été utilisées. Ces lignées ont été établies par les membres du consortium KConFab (*Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer*) (http://www.kconfab.org/Index.shtml) [334]. Les différents

défauts d'épissage ont été préalablement caractérisés par le consortium ENIGMA [322]. De plus chaque lignée a été traitée avec et sans inhibiteur de NMD. En outre les échantillons ARN de la lignée du variant *BRCA1* c.671-2A>G ont été traités par le même laboratoire afin de s'affranchir de la variabilité imputable à la culture cellulaire, le traitement inhibiteur NMD et à l'extraction de l'ARN.

A ce jour le projet QC RNA-seq regroupe 11 laboratoires dans 9 pays, dont notre laboratoire. Actuellement, 9 laboratoires ont déjà obtenu les données de séquençages RNA-seq comprenant les technologies suivantes : 3 analyses en *whole transcriptome*, 4 RNA-seq ciblés, une analyse par PCR enrichissement et une analyse par CloneSeq. CloneSeq est une technologie développée récemment par Ambry genetics [335]. Brièvement, cette technologie est basée sur le clonage de produit de PCR dans un plasmide, puis le séquençage à haut débit des transcrits issus de ces plasmides. L'analyse par PCR enrichissement a été réalisée *via* la trousse de réactifs *TruSeq Targeted RNA Expression Library Prep* (Illumina®), dont le principe est décrit en Figure 55. Au sein de ce projet QC RNA-seq, les analyses par RNA-seq ciblés varient de par la conception des sondes utilisées. Ainsi l'équipe de Rien Blok a utilisé les sondes décrites dans [335]. Le laboratoire de Petra Kleiblova a eu recours aux sondes développées pour le panel de 219 gènes CZECANCA (*CZEch CAncer paNel for Clinical Application*) [336]. L'équipe d'Anders Kvist a employé les sondes calibrées pour un panel de 19 gènes [337]. Puis enfin, notre laboratoire a utilisé les sondes de capture décrites dans [131].

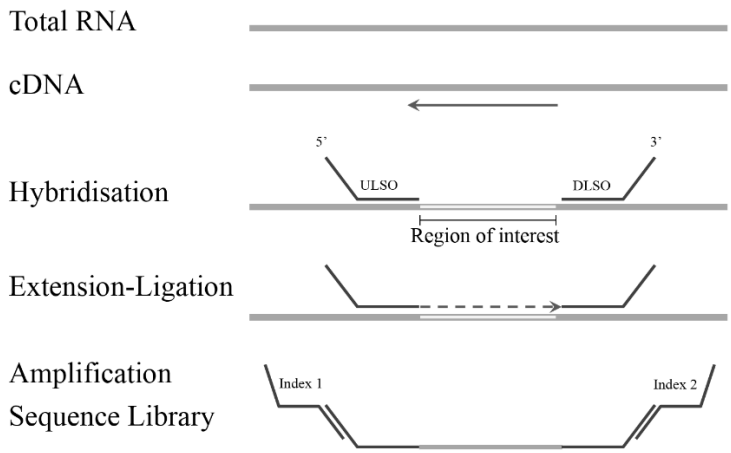


Figure 55 : Principe du TruSeq targeted RNA experiment. Après reverse transcription, les sondes s'hybrident à leur complémentaire en amont et en aval de la région d'intérêt. Une réaction d'extension-ligation joint les sondes en amont aux sondes en avales. Ces dernières sont amplifiées par PCR par des amorces ajoutant un index à chaque extrémité. Les produits de PCR sont ensuite purifiés et utilisés pour le séquençage RNA-seq.

Notre laboratoire a colligé l'ensemble des données brutes (fichier FastQ) issues des différents laboratoires. Puis nous avons appliqué le pipeline de SpliceLauncher pour extraire les informations relatives aux jonctions d'épissage. En parallèle de SpliceLauncher, nous avons aussi utilisé les outils

HTSeq count et DEXSeq [120], [141] pour évaluer la couverture des gènes *BRCA1*/*BRCA2* ainsi que de leurs exons.

Les premiers résultats, que nous avons obtenus, confirment l'intérêt du RNA-seq ciblé par rapport au RNA-seq *whole transcriptome*. En effet, une couverture plus importante des gènes *BRCA1*/*BRCA2* est observée par rapport au *whole transcriptome*. Les analyses ciblées en RNA-seq ont aussi détecté un plus large éventail de jonction d'épissage (exemple en Figure 56). En ce qui concerne l'approche par PCR enrichissement, nous avons observé une couverture inégale au sein des gènes d'intérêt. Ainsi pour certaines régions des transcrits de *BRCA1*/*BRCA2*, les épissages alternatifs n'ont pas été observés. La méthode CloneSeq a permis de détecter un grand nombre de jonction d'épissage. Cependant de par la conception de CloneSeq seuls les exons affectés par un variant ont été étudiés. Ainsi, des parties entières des transcrits *BRCA1*/*BRCA2* n'ont pas été séquencées. Par conséquent, la comparaison des données CloneSeq avec le reste des laboratoires est limitée à l'étude des variants sur l'épissage. Nous avons également constaté que la faible couverture des gènes des données de *whole transcriptome* ne permettait pas de conclure sur l'existence de jonctions d'épissage aberrantes ou renforcées par un variant. A l'inverse les données de RNA-seq ciblé ont permis de détecter des jonctions d'épissage différentiellement exprimées entre les échantillons. Les données du laboratoire d'Anders Kvist, ont permis d'identifier que l'expression des jonctions liées à des sauts d'exons (Δ9_11, Δ10_11 et Δ11) et liées à des sites accepteurs alternatifs (Δ11p(3227) et Δ10_11p(3227)) sont significativement surexprimées pour les échantillons porteurs du variant c.671-2A>G du gène *BRCA1* (Figure 57).

Splice junction	RNA isoform[a]	1	2	3	4	5	6	7	8	9
Cassette biotype *BRCA2*										
Δ2	r.-38_67del106	No	Yes	Yes	NA	Yes	No	Yes	No	No
Δ3	r.68_316del249	No	Yes	Yes	NA	Yes	Yes	Yes	Yes	Yes
Δ4	r.317_425del109	No	Yes	Yes	NA	Yes	No	Yes	No	No
Δ5	r.426_475del50	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Δ6	r.476_516del41	No	Yes	Yes	No	Yes	No	No	No	No
Δ11	r.1910_6841del4932	Yes	Yes	Yes	NA	Yes	No	Yes	No	No
Δ12	r.6842_6937del96	Yes	Yes	Yes	NA	Yes	Yes	Yes	Yes	Yes
Δ17	r.7806_7976del171	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes
Δ18	r.7977_8331del355	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Δ19	r.8332_8487del156	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Δ20	r.8488_8632del145	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Δ22	r.8755_8953del199	Yes (C+	Yes	Yes	No	Yes	Yes	Yes	No	Yes
	Total events	4	12	11	5	12	7	11	6	8

Study Site		
1	PCR enrichment	Logan/Vanessa ®
2	target RNA-seq	Petra/Jan ®
3	target RNA-seq	Rien
4	CloneSeq	Rachid ®
5	target RNA-seq	Sophie/Alexandra
6	whole transcriptome	Mads
7	target RNA-seq	Anders/Therese
8	whole transcriptome	Diana/Andrew ®
9	whole transcriptome	Inge

Figure 56 : Exemple de la détection des épissages alternatifs des sauts d'exon de *BRCA2* par SpliceLauncher. La liste des sauts d'exon correspond à celle décrite par Fackenthal [191]. Le tableau à droite assure la correspondance entre les numéros de sites et les techniques employées.

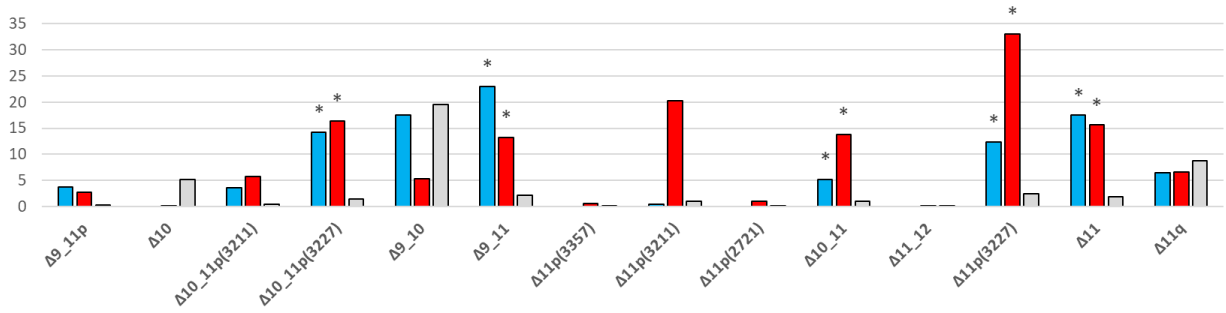


Figure 57 : Illustration des jonctions d'épissage modifiées par le variant c.671-2A>G de *BRCA1* détectés par SpliceLauncher. Les données sont issues du laboratoires d'Anders Kvist. Rectangles bleus, expression des jonctions sans inhibiteur de NMD. Rectangles rouges, expression des jonctions avec inhibiteur de NMD. Rectangles gris, moyenne d'expression des jonctions de l'ensemble des échantillons étudiés. Etoiles, expression significativement différente des autres échantillons, identifiée par SpliceLauncher.

Bien que ces premiers résultats soient particulièrement encourageants, un certain nombre de données restent à collecter ainsi que des analyses statistiques à réaliser pour compléter ces résultats. En effet, deux laboratoires n'ont pas encore séquencé les échantillons du projet QC RNA-seq. De plus ces résultats préliminaires ne permettent pas encore de proposer un seuil pour considérer un effet partiel ou total sur l'épissage par un variant. En outre d'autres laboratoires se sont proposés pour analyser les données avec leur propre pipeline. Ainsi, dans une seconde partie du projet QC RNA-seq, les pipelines d'analyses des données RNA-seq pourront être comparées entre eux. Mais cette comparaison reste à être élaborée. Ainsi, le projet QC RNA-seq permettra, au moins en partie, de proposer des recommandations pour conclure sur un défaut partiel ou total de l'épissage.

3. Un nouveau protocole de RNA-seq ciblé *long-read*

Comme a pu le montrer le projet QC RNAseq, le RNA-seq ciblé est une méthode particulièrement intéressante pour détecter des épissages alternatifs ainsi que des évènements d'épissage provoqués par un variant. Malgré tout, le RNA-seq *short read* s'avère limitant dans la compréhension de la structure complète du transcrit [119]. Si les méthodes de RNA-seq *long read* contournent cette difficulté, il n'existe pas à ce jour de protocole de capture pour le RNA-seq *long read*. Les seuls protocoles publiés de RNA-seq ciblé en long read utilisent le principe d'enrichissement par PCR. Nous pouvons citer l'utilisation de cette PCR enrichissement suivi d'un séquençage *long read* sur la plateforme Sequel de PacBio® pour identifier les épissages alternatifs de la neurexine [83]. Plus récemment, la combinaison d'une PCR d'enrichissement et de la plateforme MiniIon d'Oxford technologies® a été utilisée pour déterminer la structure des transcrits de *BRCA1* [82]. Cependant, les auteurs reconnaissent eux-mêmes la présence d'un biais majeur au niveau de la détection et de la quantification des isoformes, due à la PCR et aussi expliqué par la grande taille des transcrits pleine longueur (> 7 kb) de *BRCA1*. En marge

de ces publications, PacBio® a aussi communiqué en 2016, la possibilité d'associer leur protocole avec les sondes de capture fournies par IDT® [338].

Aussi au cours de ce travail de thèse nous avons initié la mise au point d'un protocole de capture RNA-seq en *long read* pour étudier la structure des transcrits des gènes impliqués dans le syndrome HBOC. Pour mener à bien cette étude, nous avons eu le soutien en 2018 du cancéropôle nord-ouest (CNO) (https://www.canceropole-nordouest.org/) par le biais d'un financement de 20 000 euros dans le cadre des projets émergeant du CNO. Nous avons combiné le protocole de capture fourni par Agilent® avec les sondes décrites dans l'article [131], avec le protocole de séquençage de PacBio®. Ce nouveau protocole est détaillé en annexe F. Nous l'avons appliqué sur 4 échantillons tests :

- Deux ARN issus de lymphocytes stimulés, traités ou non par un inhibiteur de NMD (puromycine).
- Un ARN extrait d'une lignée lymphoblastoïde.
  Cette lignée ainsi que les lymphocytes stimulés provenaient d'une personne indemne de toute prédisposition aux cancers du sein et de l'ovaire.
- Un ARN, extrait d'un contrôle positif : une lignée lymphoblastoïde issue d'un porteur du variant c.4096+3A>T du gène *BRCA1*.

Ces 4 échantillons ont précédemment été étudiés par RNA-seq ciblé short read. Pour l'échantillon contrôle positif, en plus des données RNA-seq, les défauts d'épissage imputables au variant c.4096+3A>T, ont préalablement été caractérisés par RT-PCR long range et multiplex. Ainsi ce variant est connu pour renforcer un site donneur alternatif (Δ11q(3309)) et le saut de l'exon 11. Nous avons sous-traité le séquençage Sequel de PacBio® par la plateforme SiRIC (SIte de Recherche Intégrée contre le Cancer) (https://siric.curie.fr/) de l'institut Curie. Pour l'analyse bioinformatique, nous avons à la fois eu recours au pipeline Iso-seq fournit par PacBio et à nos propres outils bioinformatiques.

A ce jour, un essai sans capture et un avec capture ont été réalisés. Ainsi nous avons montré la faisabilité d'une capture des transcrits des gènes impliqués dans le syndrome HBOC, avec notamment un enrichissement notable du nombre de *read* sur ces gènes lors de la capture (Tableau 14).

Tableau 14 : Différence en nombre de read après séquençage PacBio, avec ou sans capture du panel de gènes. Les gènes *BRCA1, BRCA2, PALB2, RAD51C, RAD51D* font partis de ce panel. Les transcrits du gène *ACTB* ne sont pas capturés mais le gène *ACTB* est utilisé comme gène de ménage.

Nombre de *reads*	***ACTB***	***BRCA1***	***BRCA2***	***PALB2***	***RAD51C***	***RAD51D***
Sans capture	2 986	5	4	3	3	0
Avec capture	26	277	203	139	1107	14

Pour visualiser les isoformes, les fichiers BAM obtenus après alignement sont convertis en fichiers BED. Ces derniers sont ensuite visualisés par l'outil en ligne UCSC *genome browser*. Nous pouvons constater que pour le gène *RAD51C*, à partir des données de la lignée lymphoblastoïde, certaines isoformes semblent tronquées (zones rouges) (Figure 58). Cependant un grand nombre d'entre elles supportent un transcrit complet de ce gène. En outre certains exons sont supportés par plusieurs isoformes mais non décrits dans les transcrits RefSeq (encadrés en rouges). Ainsi nous pouvons mettre en avant la capacité de cette technique à identifier de nouveaux transcrits.

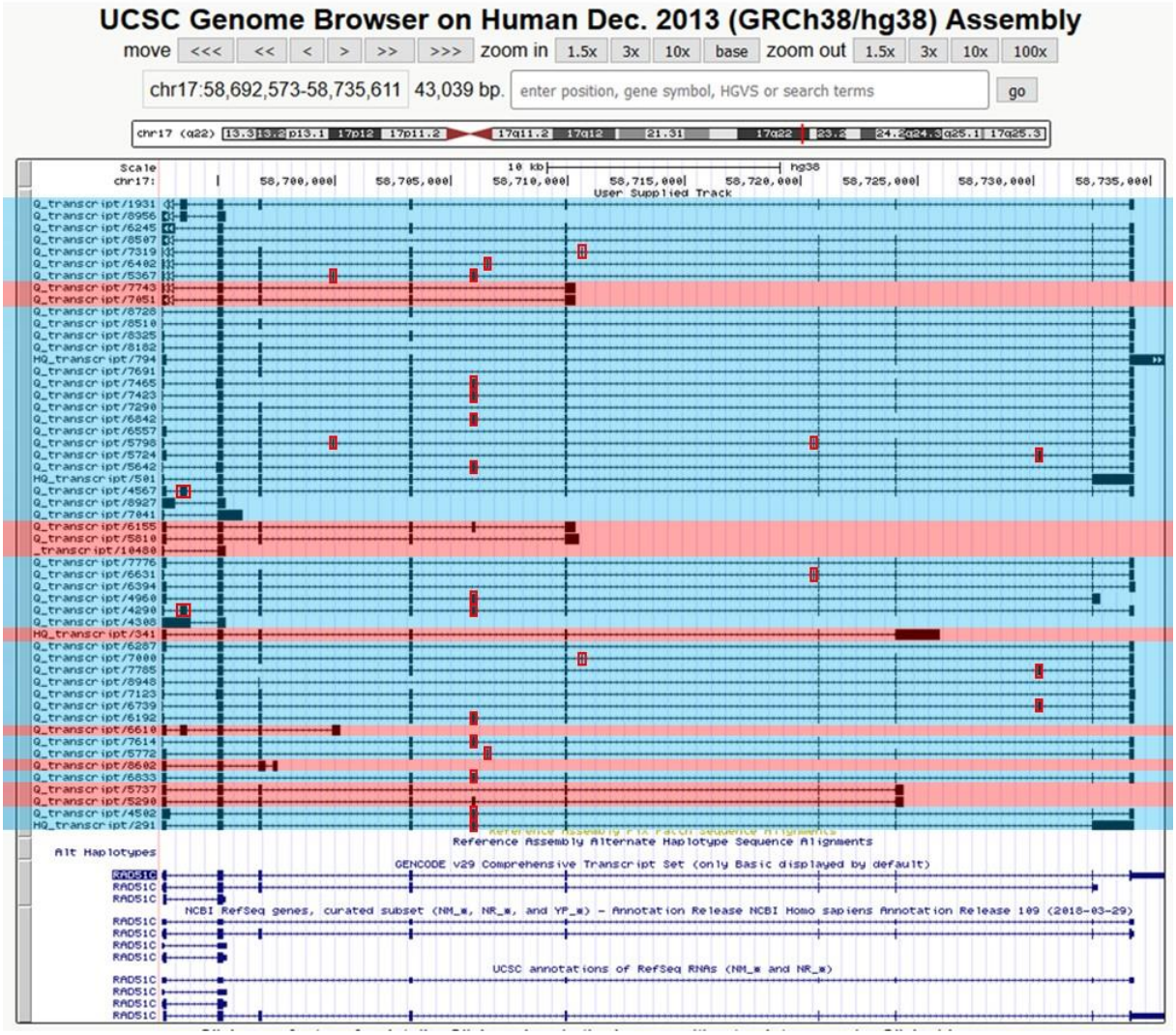


Figure 58 : Capture d'écran de l'UCSC *genome browser* pour le gène RAD51C à partir des données de la lignée lymphoblastoïde. En noir, les isoformes identifiées lors du séquençage. En bleu les transcrits décrits dans RefSeq. Les rectangles correspondent aux exons et les lignes aux introns. Les zones rouges correspondent à des exemples de transcrits tronqués. Les zones bleues correspondent à des exemples de transcrits supposés complets et les exons encadrés en rouges sont les exons non décrits dans RefSeq.

Pour la suite de ce projet, l'optimisation de ce protocole de capture est un des points à développer. En effet une analyse plus exhaustive des isoformes identifiées, nous a montré que les transcrits de grande taille (> 5 kb) étaient sous représentés. Ainsi pour le gène *BRCA1,* le transcrit complet supportant la jonction Δ11q(3309) a bien été retrouvé. Cependant pour le même échantillon, le transcrit avec l'exon 11 complet n'est pas retrouvé.

De plus les données fournies par le séquençage PacBio ne se veulent pas quantitatives. Mais pour les échantillons que nous avons séquencés, nous disposons déjà des données de RNA-seq *short read*. Aussi la suite de l'étude consistera à associer les données *long read* et *short read* pour estimer la proportion de chacune des isoformes détectées. Sachant que certains outils bioinformatiques se proposent déjà d'associer ces deux types de données. Nous pouvons ainsi citer l'outil SQANTI (*Structural and Quality Annotation of Novel Transcript Isoforms*) récemment publié, qui offre la possibilité d'évaluer la pertinence des isoformes détectées par un séquençage sur une plateforme PacBio grâce aux données *short read* [339].

4. Les forces et limites actuelles du RNA-seq pour une utilisation en diagnostic moléculaire

L'avancé du RNA-seq d'abord pour établir des profils d'expression transcriptomiques puis aujourd'hui pour l'étude de l'épissage, questionne sur l'utilisation du RNA-seq pour caractériser un défaut d'épissage imputable à un variant dans le cadre d'un diagnostic moléculaire. Les données de RNA-seq, notamment de RNA-seq ciblé, et l'émergence de nouveaux outils, tels SpliceLauncher, pour une analyse automatisée des jonctions d'épissage, offrent de nouvelles possibilités pour identifier ces évènements d'épissage. Ainsi le panel exhaustif des épissages alternatifs peut être caractérisé. Mais également l'apparition de jonctions d'épissage aberrantes ou anormalement exprimées peut également être détectée par ces nouvelles méthodes dans l'étude des variants splicéogéniques. De plus le RNA-seq ciblé *long read* complète l'étude des jonctions d'épissage par la connaissance de la structure complète des transcrits présents dans l'échantillon.

Bien que ces nouvelles technologies et méthodes améliorent l'étude des variants splicéogéniques, il persiste un certain nombre de limites pour leur applicabilité en diagnostic moléculaire. D'une part, il perdure la difficulté d'accès aux prélèvements biologiques exploitables, inhérente à toute étude ARN *in vitro*. D'autre part, force est de constater qu'actuellement nous manquons de recul pour proposer le RNA-seq afin d'aider au diagnostic moléculaire. En effet, les biologistes moléculaires ne disposent pas de recommandations pour standardiser les analyses RNA-seq ainsi que pour l'interprétation des données générées. Cette problématique a motivé le consortium ENIGMA à initier le projet QC RNA-seq. En ce qui concerne le RNA-seq *long read*, les données sont principalement qualitatives et peu, voire non quantitatives. Aussi une seconde analyse ARN *in vitro* est nécessaire pour évaluer la proportion de

chaque isoforme identifiée par RNA-seq *long read*. Or ces doubles analyses apportent une contrainte technique supplémentaire à la diffusion du RNA-seq *long read* au sein des laboratoires de diagnostic moléculaire. Néanmoins la confirmation de la proportion non négligeable de variants splicéogéniques [101] justifie l'emploi du RNA-seq dans l'étude des défauts d'épissage. Ainsi, les travaux en cours et ceux à venir pour répondre aux limites actuelles du RNA-seq sont un enjeu majeur pour aider au diagnostic moléculaire afin d'identifier de potentiels nouveaux variants pathogéniques.

Contrairement aux approches à bas débit telles que les RT-PCR, les technologies de RNA-seq identifient un grand nombre d'évènements d'épissage. De même que l'essor du NGS a entraîné la détection d'un grand nombre de variants posant ainsi la problématique de leur interprétation biologique. L'analyse par RNA-seq identifie de nombreux transcrits. Ceci soulève la question d'une part du rôle des protéines issues de leur traduction et d'autre part de leur implication dans la physiopathologie.

III. Le rôle de l'épissage dans la pathogénicité d'un variant : une histoire à suivre

Les défauts d'épissage peuvent être particulièrement complexes et sont retrouvés dans une proportion non négligeable de variants pathogéniques [46]. Si à ce jour, nous disposons de plus en plus de moyens pour prédire et caractériser ces défauts d'épissage. La question de leur impact pour l'interprétation à usage clinique de variants reste encore à être élucidée.

Une des premières étapes pour contribuer à l'interprétation des défauts d'épissage est la connaissance des épissages alternatifs. De plus, la connaissance de ces épissages alternatifs peut également jouer un rôle dans l'interprétation des variants au-delà de celle des variants splicéogéniques. Puisqu'en effet un saut d'exon, qui se produit naturellement, peut éliminer un variant non-sens et conduire à la production d'une protéine fonctionnelle [256].

Ainsi dans la cadre du syndrome HBOC, la première cartographie des épissages alternatifs des gènes *BRCA1*/*BRCA2* a été respectivement faite en 2014 et 2016 [190], [191]. Puis récemment, pour les autres principaux gènes impliqués dans le syndrome HBOC, les épissages alternatifs ont été identifiés [131], [333]. Aussi au cours de ce travail de thèse nous avons participé à l'évaluation de ces épissages alternatifs dans l'interprétation clinique des variants, en prenant comme modèle d'étude du gène *PALB2* [264]. Pour ce travail, les épissages alternatifs de ce gène ont été caractérisés à partir d'un panel d'échantillons représentatifs des tissus sanguins, mammaires et ovariens. Puis, pour chacun de ces épissages, leur capacité à sauvegarder la fonctionnalité de PALB2, malgré la présence de variants *a priori* délétères, a été évaluée. Il a ainsi été identifié 88 événements d'épissage. Pour les variants non-sens, nous n'avons pas identifié de transcrits susceptibles de sauvegarder la fonctionnalité de la protéine. Effectivement, soit ces transcrits étaient supposés fonctionnels mais trop faiblement exprimés (< 1 % du transcrit de référence), soit nettement exprimés (8-34 %) mais supposés non fonctionnels. En effet, le saut de l'exon 12 a été observé jusqu'à 34 % du transcrit naturel mais il emporte le domaine WD40. De plus, le variant c.3201+5G>T a été récemment publié comme entrainant un renforcement majeur des sauts de l'exon 12 et 11-12 [340]. Les tests fonctionnels chez les porteurs de ce variant semblent confirmer la perte de fonction de la protéine PALB2. Cependant pour les variants situés sur les motifs canoniques accepteurs (AG) des exons 2, 5, 7, et 10 ; leurs effets semblent être tempérés par la présence de sites accepteurs alternatifs. Ces sites accepteurs conduisent en effet à la génération de transcrits supposés fonctionnels. Ces données peuvent d'ores et déjà suggérer une possible réévaluation de 43 variants rapportés comme pathogènes ou probablement pathogènes dans la base de données ClinVar.

Par ailleurs, l'exon alternatif, IVS1-463▼(134), est décrit comme faiblement exprimé dans l'intron 1 de *PALB2* (3 à 5 % du transcrit de référence) [264]. Cependant il a été récemment montré que le variant c.108+1G>A situé sur le site donneur de l'exon 2 entraine le saut de l'exon 2 mais aussi le renforcement significatif de cet exon alternatif (données non publiées) (Figure 59). Cette étude a été réalisée dans le

cadre du réseau épissage du GCC. Ainsi nous pouvons conclure que même un épissage alternatif complexe et à distance d'un variant peut-être renforcé par l'action de ce variant.

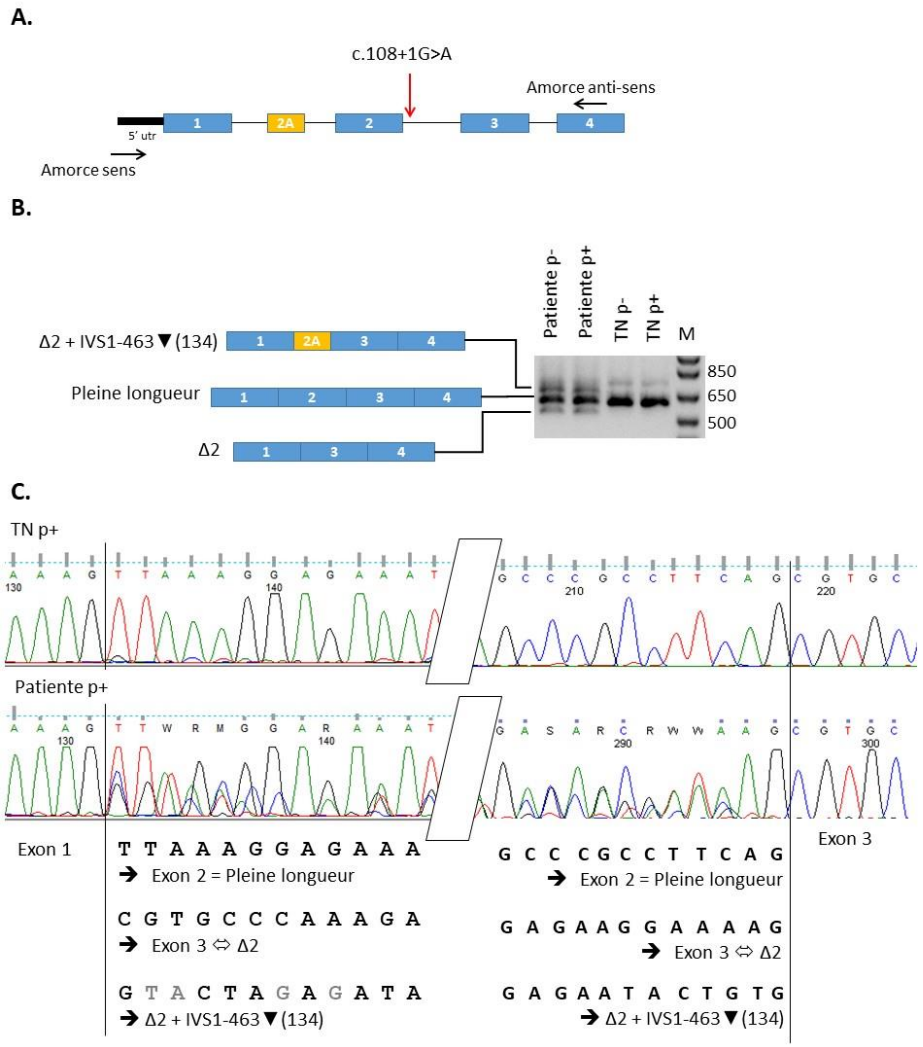


Figure 59 : Etude ARN *in vitro* du variant c.108+1G>A du gène *PALB2*. Utilisation d'ARN de lignées lymphoblastoïdes. **A.** Schéma du pré-ARNm avec la position du variant c.108+1G>A et position des amorces dans le 5' UTR et dans l'exon 4, l'exon alternatif IVS1-463▼(134) est noté 2A. **B.** Résultat de la RT-PCR avec la description des bandes observées. **C.** Séquençage Sanger des produits de la RT-PCR. TN : témoin négatif, M : marqueur de taille, P+ : avec inhibiteur de NMD (puromycine), P- : sans inhibiteur de NMD (puromycine)

Ainsi la classification systématique des variants comme pathogènes ou probablement pathogènes se doit d'être pondérée par la présence des épissages alternatifs. En effet, dans le modèle du gène PALB2, bien que la plupart des épissages alternatifs soient faiblement exprimés au regard du transcrit de référence, ils sont suffisants pour amener à reconsidérer la pathogénicité de certains variants canoniques. Or, pour le syndrome HBOC, de plus en plus de gènes y sont associés [341]. De même pour les paralogues du gène *RAD51* (*RAD51C/D*) qui ont été récemment inclus dans le panel de gènes testés pour le diagnostic moléculaire du syndrome HBOC [215], [216]. En outre les protéines RAD51C/D ont été décrits comme

présentant des épissages alternatifs majeurs [131], [333]. Pour *RAD51C*, si les épissages alternatifs sont nombreux, ils sont faiblement exprimés (< 10 % du transcrit de référence). Alors que les transcrits de *RAD51D* présentent d'important épissages alternatifs. A titre d'exemple, le saut des exons 3 est fortement exprimé. En effet, le nombre de transcrits est égale voire supérieur au nombre de transcrits de référence [131]. De plus ces épissages alternatifs semblent être tissu dépendant et modifier la fonctionnalité de la protéine [342], [343]. Aussi la question de la sauvegarde de la fonctionnalité de la protéine malgré la présence de variant délétère se pose également pour ce nouveau gène. En effet, déjà 9 variants rapportés comme pathogènes par la base de données ClinVar sont situés dans l'exon 3 de *RAD51D* (octobre 2019). Cependant il apparait que des études complémentaires soient nécessaires pour pouvoir confirmer leurs significations cliniques.

En addition à ces nouveaux gènes, il existe toujours certaines questions sur la fonctionnalité des épissages alternatifs de *BRCA1/BRCA2*. Alors que les variants de ces deux gènes sont étudiés depuis plus de 20 ans. Parmi les points troubles de l'interprétation à usage clinique, l'épissage alternatif de l'exon 11 de *BRCA1* soulève un certain nombre d'interrogations. Physiologiquement, cet exon subit une troncation de 3 309 nt, par l'utilisation d'un site donneur alternatif (Δ11q(3309)). Cet épissage est observé dans environ 10 % des transcrits. Bien que cette délétion ne change pas le cadre de lecture et n'affecte pas les domaines fonctionnels majeurs (RING et BRCT), elle induit une perte de 59.2 % de la séquence codante. La compréhension de la fonctionnalité des protéines produites par ce transcrit se révèle être cruciale. C'est pourquoi plusieurs études ont été menées pour apporter des éléments de réponse sur cette fonctionnalité. Il en résulte que ce transcrit Δ11q(3309) semble en partie maintenir la fonctionnalité de BRCA1 [344]–[346]. Ainsi donc la pathogénicité des variants situés dans cette région semble devoir être pondérée par cet épissage alternatif. De même que les variants altérant le site donneur naturel de l'exon, comme le variant c.4096+3A>G, conduisent majoritairement à la production du transcrit Δ11q(3309) [347]. Or le variant c.4096+3A>G de *BRCA1* a été récemment identifié chez un porteur sain à l'état homozygote [348]. Tandis que la présence d'un variant délétère homozygote de *BRCA1* est considéré comme n'étant pas viable au stade embryologique. A l'inverse le consortium ENIGMA a identifié le variant c.4096+1G>A. Les données cliniques et familiales tendent à considérer ce variant comme pathogène. En effet, les données ont été collectées à partir de 28 familles porteuses du variant c.4096+3A>G et de 18 familles porteuses du variant c.4096+1G>A (données non publiées, communication ENIGMA, Avril 2019, USA). La probabilité *a posteriori* que le variant c.4096+3A>G soit pathogène est de 0.000034, soit classe 1 (variant bénin). Alors que pour le variant c.4096+1G>A, cette probabilité est de 0.99999, soit classe 5 (variant pathogène). Ce qui est d'autant plus troublant que lorsque nous avons étudié l'ARN de patients porteurs du variant c.4096+3A>G et du variant c.4096+1G>A, les défauts d'épissage semblent être similaires (données non publiées) (Figure 60).

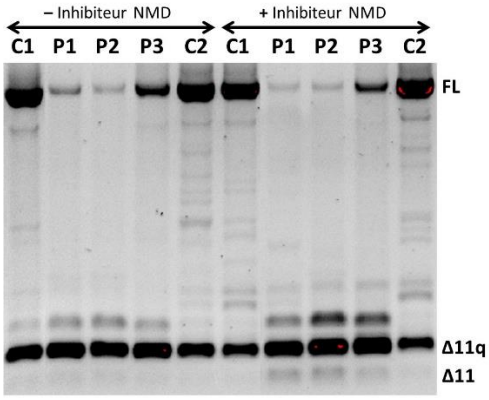


Figure 60 : Comparaison des défauts d'épissage de l'exon 11 par RT-PCR long range pour les variants BRCA1 c.4096+1G>A, c.4096+3A>G et c.4096+3A>T. Inhibiteur NMD : puromycine ; NMD : Nonsense-Mediated Decay ; C1 et C2 : contrôle 1 et 2 ; P1 : porteur du variant c.4096+1G>A ; P2 : porteur du variant c.4096+3A>G ; P3 : porteur du variant c.4096+3A>T ; FL : full length ; Δ11q : perte des 3 309 nt de l'exon 11, Δ11q(3309) ; Δ11 : saut de l'exon 11.

Face à ces résultats contradictoires, le consortium ENIGMA a initié un nouveau projet pour collecter les données fonctionnelles, cliniques et familiales des variants situés sur les sites canoniques d'épissage des gènes *BRCA1*/*BRCA2*. Ce projet codirigé par l'équipe de Miguel de la Hoya et notre laboratoire, permettra ainsi de proposer une interprétation clinique de ces variants.

Dès lors nous pouvons dire que le renforcement des épissages alternatifs par un variant splicéogénique ne garantit pas l'absence d'impact sur la fonctionnalité de la protéine. En effet, la Δ11q(3309) de *BRCA1*, le saut de l'exon 12 de *PALB2* ou bien le saut de l'exon 3 de *BRCA2* ont tous en commun d'être des épissages alternatifs majeurs de ces gènes [131], [190], [191], [264]. Cependant, les données cliniques et fonctionnelles tendent à montrer que les protéines traduites sont partiellement voire non fonctionnelles [251], [340]. Aussi la question de savoir si ces transcrits et les protéines résultantes peuvent avoir un rôle à jouer dans la cellule ou seulement représenter des erreurs systématiques de l'épissage reste entière. D'autant plus qu'à l'échelle de l'évolution des eucaryotes, l'épissage alternatif est supposé avoir eu un rôle conséquent [31]. Aussi les transcrits puis les protéines produites pas ces épissages ne peuvent être considérés, au moins en partie, que comme des sous-produits aléatoires et sans impact sur l'organisme [349].

Que ce soit par la diversité des modifications d'épissage ou bien par la compréhension du rôle de ces épissages, le lien entre un variant splicéogénique et la pathogénicité du dit variant est complexe et versatile. Cependant, plusieurs études se sont proposées d'associer directement la prédiction d'un variant splicéogénique à la pathogénicité. Ainsi nous pouvons citer le travail publié en 2016 de l'équipe de Sean Tavtigian. Ce travail avait pour but d'estimer en fonction du score MES, transformé en z-score, la probabilité *a priori* que le variant soit pathogène pour les gènes *BRCA1*/*BRCA2* [268]. En effet cette équipe avait déjà utilisé une approche similaire pour les variants faux sens avec le score AGVGD [225]. Néanmoins l'utilisation du score MES ne permet pas d'étudier les variants impactant les points de

branchement ou les motifs régulateurs comme les ESRs. De plus pour les variants créant un nouveau site d'épissage, les auteurs n'avaient pas suffisamment de données pour calculer une probabilité *a priori* d'être pathogène. Ainsi les auteurs ont pu estimer cette probabilité seulement pour les variants situés dans les motifs consensus des sites d'épissage donneur/accepteur. De plus les auteurs reconnaissent que, même pour ces variants consensus, plusieurs éléments en plus du score MES doivent être pris en compte pour évaluer la probabilité *a priori* d'être pathogènes.

Plus récemment, nous pouvons aussi citer l'outil MMSplice publié en 2019 par l'équipe de Julien Gagneur [293]. Cet outil a été évalué pour identifier des variants pathogéniques issus de la base de données ClinVar. Ainsi, MMSplice a montré une performance optimale pour discriminer les variants neutres ou pathogènes (AUC = 0.95). Néanmoins, d'une part les auteurs ne proposent pas de probabilités *a priori* d'être pathogène pour un modèle multifactoriel ou de classe d'argument compatible avec la classification ACMG. D'autre part, les variants utilisés pour cette évaluation étaient situés dans ou proche des motifs consensus d'épissage donneur/accepteur (-10/+10 et -50/+10). Or il a depuis peu été montré que la majorité des variants splicéogéniques sont situés en dehors de ces sites d'épissage [101].

Il en résulte donc que les prédictions de pathogénicité ne sont actuellement utilisables que pour les variants dans les sites consensus d'épissage. Par ailleurs, dans le cadre du diagnostic moléculaire, elles ne sont pas suffisantes pour calculer la probabilité de pathogénicité, comme a pu le montrer le z-score de MES [268]. Aussi ces prédictions sont intéressantes au titre de la recherche, mais pour le diagnostic moléculaire, il persiste la nécessité de caractériser *in vitro* les modifications d'épissage. De ce fait, l'application d'analyse ARN à haut débit, telle le RNA-seq, apparaît comme de plus en plus pertinente pour le diagnostic moléculaire.

Depuis le milieu des années 2010, il est apparu un élan pour élargir les tests génétiques dans le cadre du syndrome HOBC, notamment soutenu par le Dr Marie-Claire King qui a prouvé le lien entre les variants de *BRCA1* et le syndrome HBOC [350]. En plus d'élargir le panel de gènes à tester pour le syndrome HBOC, il est proposé de rechercher des variants pathogènes indépendamment des histoires familiales et des données cliniques. En effet, de récents travaux ont révélé qu'environ la moitié des personnes porteuses de variants pathogènes ne remplissaient pas les critères de recrutement issus des recommandations pour une orientation vers un test génétique [351]. Néanmoins pour ces nouveaux variants, il n'existe pas de données cliniques ou fonctionnelles pour confirmer leur pathogénicité. Mais si ces résultats sont confirmés, alors le nombre de personnes éligibles au diagnostic moléculaire du syndrome HBOC va croître de façon exponentielle. En effet la proposition de Marie-Claire King est de tester toutes les femmes atteignant 30 ans. Juste en France, le nombre annuel de personnes testées augmenterait de 20 000 à près de 400 000 (données INSEE 2018). Aussi face à cette possible masse de variants à venir, l'interprétation à usage clinique des variants se devra d'atteindre un niveau de précision inédit pour demeurer cliniquement et éthiquement pertinente.

Compte tenu de la proportion de variants splicéogéniques, la caractérisation et l'interprétation des modifications de l'épissage sont des points cruciaux pour répondre aux défis présents et futurs du diagnostic moléculaire.

REFERENCES

[1] Y. Abel, G. Clerget, V. Bourguignon-Igel, V. Salone, and M. Rederstorff, 'Les petits ARN nucléolaires nous surprennent encore !', *médecine/sciences*, vol. 30, no. 3, pp. 297–302, Mar. 2014.

[2] S. Baulande, A. Criqui, and M. Duthieuw, 'Les microARN circulants, une nouvelle classe de biomarqueurs pour la médecine', *médecine/sciences*, vol. 30, no. 3, pp. 289–296, Mar. 2014.

[3] T. Pedrazzini, 'Le cœur des ARN non codants - Un long chemin à découvrir', *médecine/sciences*, vol. 31, no. 3, pp. 261–267, Mar. 2015.

[4] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, 'An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA', *Cell*, vol. 12, no. 1, pp. 1–8, Sep. 1977.

[5] W. Gilbert, 'Why genes in pieces?', *Nature*, vol. 271, no. 5645, p. 501, Feb. 1978.

[6] N. A. O'Leary *et al.*, 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, Jan. 2016.

[7] M. C. Wahl, C. L. Will, and R. Lührmann, 'The Spliceosome: Design Principles of a Dynamic RNP Machine', *Cell*, vol. 136, no. 4, pp. 701–718, Feb. 2009.

[8] R. Reed, 'Initial splice-site recognition and pairing during pre-mRNA splicing', *Curr. Opin. Genet. Dev.*, vol. 6, no. 2, pp. 215–220, Apr. 1996.

[9] N. Sheth, X. Roca, M. L. Hastings, T. Roeder, A. R. Krainer, and R. Sachidanandam, 'Comprehensive splice-site analysis using comparative genomics', *Nucleic Acids Res.*, vol. 34, no. 14, pp. 3955–3967, Sep. 2006.

[10] A. A. Patel and J. A. Steitz, 'Splicing double: insights from the second spliceosome', *Nat. Rev. Mol. Cell Biol.*, vol. 4, no. 12, p. 960, Dec. 2003.

[11] A. Levine and R. Durbin, 'A computational scan for U12-dependent introns in the human genome sequence', *Nucleic Acids Res.*, vol. 29, no. 19, pp. 4006–4013, Oct. 2001.

[12] R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, and P. Chambon, 'Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries', *Proc. Natl. Acad. Sci.*, vol. 75, no. 10, pp. 4853–4857, Oct. 1978.

[13] C. B. Burge, T. Tuschi, and P. A. Sharp, 'Splicing of Precursors to mRNAs by the Spliceosomes', in *The RNA World II*, Cold Spring Harbor Laboratory Press, 1999, pp. 525–560.

[14] M. Li and P. H. Pritchard, 'Characterization of the Effects of Mutations in the Putative Branchpoint Sequence of Intron 4 on the Splicing within the Human Lecithin:cholesterol Acyltransferase Gene', *J. Biol. Chem.*, vol. 275, no. 24, pp. 18079–18084, Jun. 2000.

[15] T. R. Mercer *et al.*, 'Genome-wide discovery of human splicing branchpoints', *Genome Res.*, p. gr.182899.114, Jan. 2015.

[16] H. Sun and L. A. Chasin, 'Multiple Splicing Defects in an Intronic False Exon', *Mol. Cell. Biol.*, vol. 20, no. 17, pp. 6414–6425, Sep. 2000.

[17] B. R. Graveley, 'Sorting out the complexity of SR protein functions', *RNA*, vol. 6, no. 9, pp. 1197–1211, Sep. 2000.

[18] U. Pozzoli and M. Sironi, 'Silencers regulate both constitutive and alternative splicing events in mammals', *Cell. Mol. Life Sci. CMLS*, vol. 62, no. 14, pp. 1579–1604, Jul. 2005.

[19] W. G. Fairbrother, R.-F. Yeh, P. A. Sharp, and C. B. Burge, 'Predictive Identification of Exonic Splicing Enhancers in Human Genes', *Science*, vol. 297, no. 5583, pp. 1007–1013, Aug. 2002.

[20] Z. Wang, M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge, 'Systematic Identification and Analysis of Exonic Splicing Silencers', *Cell*, vol. 119, no. 6, pp. 831–845, Dec. 2004.

[21] P. J. Smith, C. Zhang, J. Wang, S. L. Chew, M. Q. Zhang, and A. R. Krainer, 'An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers', *Hum. Mol. Genet.*, vol. 15, no. 16, pp. 2490–2508, Aug. 2006.

[22] A. Goren *et al.*, 'Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers', *Mol. Cell*, vol. 22, no. 6, pp. 769–781, Jun. 2006.

[23] S. Ke, X. H.-F. Zhang, and L. A. Chasin, 'Positive selection acting on splicing motifs reflects compensatory evolution', *Genome Res.*, vol. 18, no. 4, pp. 533–543, Jan. 2008.
[24] S. Ke *et al.*, 'Quantitative evaluation of all hexamers as exonic splicing elements', *Genome Res.*, vol. 21, no. 8, pp. 1360–1374, Jan. 2011.
[25] A. B. Rosenberg, R. P. Patwardhan, J. Shendure, and G. Seelig, 'Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences', *Cell*, vol. 163, no. 3, pp. 698–711, Oct. 2015.
[26] Z. Wang and C. B. Burge, 'Splicing regulation: From a parts list of regulatory elements to an integrated splicing code', *RNA*, vol. 14, no. 5, pp. 802–813, Jan. 2008.
[27] D. Schmucker *et al.*, 'Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity', *Cell*, vol. 101, no. 6, pp. 671–684, Jun. 2000.
[28] D. S. Navaratnam, T. J. Bell, T. D. Tu, E. L. Cohen, and J. C. Oberholtzer, 'Differential Distribution of Ca2+-Activated K+ Channel Splice Variants among Hair Cells along the Tonotopic Axis of the Chick Cochlea', *Neuron*, vol. 19, no. 5, pp. 1077–1085, Nov. 1997.
[29] K. P. Rosenblatt, Z.-P. Sun, S. Heller, and A. J. Hudspeth, 'Distribution of Ca2+-Activated K+ Channel Isoforms along the Tonotopic Gradient of the Chicken's Cochlea', *Neuron*, vol. 19, no. 5, pp. 1061–1075, Nov. 1997.
[30] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat. Genet.*, vol. 40, no. 12, pp. 1413–1415, Dec. 2008.
[31] N. L. Barbosa-Morais *et al.*, 'The Evolutionary Landscape of Alternative Splicing in Vertebrate Species', *Science*, vol. 338, no. 6114, pp. 1587–1593, Dec. 2012.
[32] L. H. Boise *et al.*, 'bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death', *Cell*, vol. 74, no. 4, pp. 597–608, Aug. 1993.
[33] L. Sánchez, 'Sex-determining mechanisms in insects', *Int. J. Dev. Biol.*, vol. 52, no. 7, pp. 837–856, Sep. 2004.
[34] E. T. Wang *et al.*, 'Alternative isoform regulation in human tissue transcriptomes', *Nature*, vol. 456, no. 7221, pp. 470–476, Nov. 2008.
[35] K. W. Lynch, 'Regulation of alternative splicing by signal transduction pathways.', *Adv. Exp. Med. Biol.*, vol. 623, pp. 161–174, 2007.
[36] H. Jung *et al.*, 'Intron retention is a widespread mechanism of tumor-suppressor inactivation', *Nat. Genet.*, vol. 47, no. 11, pp. 1242–1248, Nov. 2015.
[37] T. Ø. Tange, A. Nott, and M. J. Moore, 'The ever-increasing complexities of the exon junction complex', *Curr. Opin. Cell Biol.*, vol. 16, no. 3, pp. 279–284, Jun. 2004.
[38] E. Nagy and L. E. Maquat, 'A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance', *Trends Biochem. Sci.*, vol. 23, no. 6, pp. 198–199, Jun. 1998.
[39] B. P. Lewis, R. E. Green, and S. E. Brenner, 'Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans', *Proc. Natl. Acad. Sci.*, vol. 100, no. 1, pp. 189–192, Jan. 2003.
[40] M. W. Popp and L. E. Maquat, 'Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine', *Cell*, vol. 165, no. 6, pp. 1319–1322, Jun. 2016.
[41] E. D. Karousis, S. Nasif, and O. Mühlemann, 'Nonsense-mediated mRNA decay: novel mechanistic insights and biological impact', *Wiley Interdiscip. Rev. RNA*, vol. 7, no. 5, pp. 661–682, 2016.
[42] Y.-F. Chang, J. S. Imam, and M. F. Wilkinson, 'The Nonsense-Mediated Decay RNA Surveillance Pathway', *Annu. Rev. Biochem.*, vol. 76, no. 1, pp. 51–74, 2007.
[43] S. H. Orkin *et al.*, 'Linkage of β -thalassaemia mutations and β -globin gene polymorphisms with DNA polymorphisms in human β -globin gene cluster', *Nature*, vol. 296, no. 5858, p. 627, Apr. 1982.
[44] R. K. Singh and T. A. Cooper, 'Pre-mRNA splicing in disease and therapeutics', *Trends Mol. Med.*, vol. 18, no. 8, pp. 472–482, Aug. 2012.
[45] A. Kalsotra and T. A. Cooper, 'Functional consequences of developmentally regulated alternative splicing', *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 715–729, Oct. 2011.

[46] N. López-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigó, 'Are splicing mutations the most frequent cause of hereditary disease?', *FEBS Lett.*, vol. 579, no. 9, pp. 1900–1903, 2005.
[47] M. Krawczak *et al.*, 'Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mRNA splicing', *Hum. Mutat.*, vol. 28, no. 2, pp. 150–158, 2007.
[48] K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, and W. G. Fairbrother, 'Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes', *Proc. Natl. Acad. Sci.*, vol. 108, no. 27, pp. 11093–11098, Jul. 2011.
[49] T. Sterne-Weiler, J. Howard, M. Mort, D. N. Cooper, and J. R. Sanford, 'Loss of exon identity is a common mechanism of human inherited disease', *Genome Res.*, vol. 21, no. 10, pp. 1563–1571, Jan. 2011.
[50] X. Roca, R. Sachidanandam, and A. R. Krainer, 'Intrinsic differences between authentic and cryptic 5′ splice sites', *Nucleic Acids Res.*, vol. 31, no. 21, pp. 6321–6333, Nov. 2003.
[51] A. Anna and G. Monika, 'Splicing mutations in human genetic disorders: examples, detection, and confirmation', *J. Appl. Genet.*, vol. 59, no. 3, pp. 253–268, Aug. 2018.
[52] M. Miné, M. Brivet, G. Touati, P. Grabowski, M. Abitbol, and C. Marsac, 'Splicing Error in E1α Pyruvate Dehydrogenase mRNA Caused by Novel Intronic Mutation Responsible for Lactic Acidosis and Mental Retardation', *J. Biol. Chem.*, vol. 278, no. 14, pp. 11768–11772, Apr. 2003.
[53] K. Wimmer *et al.*, 'Extensive in silico analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5′ splice-site disruption', *Hum. Mutat.*, vol. 28, no. 6, pp. 599–612, Jun. 2007.
[54] C. Dobkin, R. G. Pergolizzi, P. Bahre, and A. Bank, 'Abnormal splice in a mutant human beta-globin gene not at the site of a mutation', *Proc. Natl. Acad. Sci.*, vol. 80, no. 5, pp. 1184–1188, Mar. 1983.
[55] R. Vaz-Drago, N. Custódio, and M. Carmo-Fonseca, 'Deep intronic mutations and human disease', *Hum. Genet.*, vol. 136, no. 9, pp. 1093–1111, Sep. 2017.
[56] N. Deburgrave *et al.*, 'Protein- and mRNA-based phenotype–genotype correlations in DMD/BMD with point mutations and molecular basis for BMD with nonsense and frameshift mutations in the DMD gene', *Hum. Mutat.*, vol. 28, no. 2, pp. 183–195, 2007.
[57] J. C. Alwine, D. J. Kemp, and G. R. Stark, 'Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5350–5354, 1977.
[58] F. C. Kafatos, C. W. Jones, and A. Efstratiadis, 'Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure', *Nucleic Acids Res.*, vol. 7, no. 6, pp. 1541–1552, Nov. 1979.
[59] R. K. Saiki *et al.*, 'Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia', *Science*, vol. 230, no. 4732, pp. 1350–1354, Dec. 1985.
[60] H. M. Temin, 'RNA-directed DNA synthesis', *Sci. Am.*, vol. 226, no. 1, pp. 25–33, Jan. 1972.
[61] C. Orlando, P. Pinzani, and M. Pazzagli, 'Developments in quantitative PCR.', *Clin. Chem. Lab. Med.*, vol. 36, no. 5, pp. 255–269, May 1998.
[62] S. Cheng, C. Fockler, W. M. Barnes, and R. Higuchi, 'Effective amplification of long targets from cloned inserts and human genomic DNA', *Proc. Natl. Acad. Sci.*, vol. 91, no. 12, pp. 5695–5699, Jun. 1994.
[63] M. Claustres *et al.*, '[Detection of deletions by the amplification of exons (multiplex PCR) in Duchenne muscular dystrophy].', *J. Genet. Hum.*, vol. 37, no. 3, pp. 251–257, Sep. 1989.
[64] P.-L. Quan, M. Sauzade, and E. Brouzes, 'dPCR: A Technology Review', *Sensors*, vol. 18, no. 4, p. 1271, Apr. 2018.
[65] C. M. Hindson *et al.*, 'Absolute quantification by droplet digital PCR versus analog real-time PCR', *Nat. Methods*, vol. 10, no. 10, pp. 1003–1005, Oct. 2013.
[66] M. A. Jenkins and M. D. Guerin, 'Capillary electrophoresis as a clinical tool', *J. Chromatogr. B. Biomed. Sci. App.*, vol. 682, no. 1, pp. 23–34, Jun. 1996.
[67] F. Sanger, S. Nicklen, and A. R. Coulson, 'DNA sequencing with chain-terminating inhibitors', *Proc. Natl. Acad. Sci.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977.

[68] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, 'Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray', *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995.
[69] C. Sotiriou and L. Pusztai, 'Gene-Expression Signatures in Breast Cancer', *N. Engl. J. Med.*, vol. 360, no. 8, pp. 790–800, Feb. 2009.
[70] E. R. Mardis, 'Next-Generation DNA Sequencing Methods', *Annu. Rev. Genomics Hum. Genet.*, vol. 9, no. 1, pp. 387–402, 2008.
[71] H. P. J. Buermans and J. T. den Dunnen, 'Next generation sequencing technology: Advances and applications', *Biochim. Biophys. Acta BBA - Mol. Basis Dis.*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014.
[72] F. Ozsolak *et al.*, 'Direct RNA sequencing', *Nature*, vol. 461, no. 7265, pp. 814–818, Oct. 2009.
[73] O. Stegle, S. A. Teichmann, and J. C. Marioni, 'Computational and analytical challenges in single-cell transcriptomics', *Nat. Rev. Genet.*, vol. 16, no. 3, pp. 133–145, Mar. 2015.
[74] H. Ozcelik *et al.*, 'Long-Range PCR and Next-Generation Sequencing of BRCA1 and BRCA2 in Breast Cancer', *J. Mol. Diagn.*, vol. 14, no. 5, pp. 467–475, Sep. 2012.
[75] E. Samorodnitsky *et al.*, 'Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing', *Hum. Mutat.*, vol. 36, no. 9, pp. 903–914, 2015.
[76] J. Z. Levin *et al.*, 'Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts', *Genome Biol.*, vol. 10, no. 10, p. R115, Oct. 2009.
[77] T. R. Mercer *et al.*, 'Targeted RNA sequencing reveals the deep complexity of the human transcriptome', *Nat. Biotechnol.*, vol. 30, no. 1, pp. 99–104, Jan. 2012.
[78] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, 'Continuous base identification for single-molecule nanopore DNA sequencing', *Nat. Nanotechnol.*, vol. 4, no. 4, pp. 265–270, Apr. 2009.
[79] J. Eid *et al.*, 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science*, vol. 323, no. 5910, pp. 133–138, Jan. 2009.
[80] M. T. Bolisetty, G. Rajadinakaran, and B. R. Graveley, 'Determining exon connectivity in complex mRNAs by nanopore sequencing', *Genome Biol.*, vol. 16, no. 1, p. 204, Sep. 2015.
[81] A. Rhoads and K. F. Au, 'PacBio Sequencing and Its Applications', *Genomics Proteomics Bioinformatics*, vol. 13, no. 5, pp. 278–289, Oct. 2015.
[82] L. C. de Jong *et al.*, 'Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events', *Breast Cancer Res.*, vol. 19, no. 1, p. 127, Nov. 2017.
[83] B. Treutlein, O. Gokce, S. R. Quake, and T. C. Südhof, 'Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing', *Proc. Natl. Acad. Sci.*, vol. 111, no. 13, pp. E1291–E1299, Apr. 2014.
[84] A. Schroeder *et al.*, 'The RIN: an RNA integrity number for assigning integrity values to RNA measurements', *BMC Mol. Biol.*, vol. 7, no. 1, p. 3, Jan. 2006.
[85] O. Mueller *et al.*, 'A microfluidic system for high-speed reproducible DNA sizing and quantitation', *ELECTROPHORESIS*, vol. 21, no. 1, pp. 128–134, 2000.
[86] L. Rainen *et al.*, 'Stabilization of mRNA Expression in Whole Blood Samples', *Clin. Chem.*, vol. 48, no. 11, pp. 1883–1890, Nov. 2002.
[87] C. Williams *et al.*, 'A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens', *Am. J. Pathol.*, vol. 155, no. 5, pp. 1467–1471, Nov. 1999.
[88] R. Martin, A. E. Mogg, L. A. Heywood, L. Nitschke, and J. F. Burke, 'Aminoglycoside suppression at UAG, UAA and UGA codons inEscherichia coli and human tissue culture cells', *Mol. Gen. Genet. MGG*, vol. 217, no. 2, pp. 411–418, Jun. 1989.
[89] J. F. Burke and A. E. Mogg, 'Suppression of a nonsense mutation in mammalian cells in vivo by the aminoglycoside anthiotics G–418 and paromomycin', *Nucleic Acids Res.*, vol. 13, no. 17, pp. 6265–6272, Sep. 1985.
[90] L. Perrin-Vidoz, O. M. Sinilnikova, D. Stoppa-Lyonnet, G. M. Lenoir, and S. Mazoyer, 'The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons', *Hum. Mol. Genet.*, vol. 11, no. 23, pp. 2805–2814, Nov. 2002.

[91] J. Sambrook, E. F. Fritsch, and T. Maniatis, 'Molecular cloning: a laboratory manual.', *Mol. Cloning Lab. Man.*, no. Ed. 2, 1989.
[92] P. Gaildrat *et al.*, 'Multiple sequence variants of BRCA2 exon 7 alter splicing regulation', *J. Med. Genet.*, vol. 49, no. 10, pp. 609–617, Oct. 2012.
[93] A. Acedo *et al.*, 'Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes', *Breast Cancer Res.*, vol. 14, no. 3, p. R87, May 2012.
[94] A. Y. Steffensen *et al.*, 'Functional characterization of *BRCA1* gene variants by mini-gene splicing assay', *Eur. J. Hum. Genet.*, vol. 22, no. 12, pp. 1362–1368, Dec. 2014.
[95] C. Bonnet *et al.*, 'Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene', *J. Med. Genet.*, vol. 45, no. 7, pp. 438–446, Jul. 2008.
[96] H. M. van der Klift *et al.*, 'Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses', *Mol. Genet. Genomic Med.*, vol. 3, no. 4, pp. 327–345, Jul. 2015.
[97] P. Gaildrat, A. Killian, A. Martins, I. Tournier, T. Frébourg, and M. Tosi, 'Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants.', *Methods Mol. Biol. Clifton NJ*, vol. 653, pp. 249–257, 2010.
[98] A. Melnikov *et al.*, 'Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay', *Nat. Biotechnol.*, vol. 30, no. 3, pp. 271–277, Mar. 2012.
[99] R. Soemedi *et al.*, 'Pathogenic variants that alter protein code often disrupt splicing', *Nat. Genet.*, vol. 49, no. 6, p. 848, Jun. 2017.
[100] S. I. Adamson, L. Zhan, and B. R. Graveley, 'Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency', *Genome Biol.*, vol. 19, no. 1, p. 71, Jun. 2018.
[101] R. Cheung *et al.*, 'A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions', *Mol. Cell*, vol. 73, no. 1, pp. 183-194.e8, Jan. 2019.
[102] M. Lek *et al.*, 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, vol. 536, no. 7616, pp. 285–291, Aug. 2016.
[103] W. R. Pearson and D. J. Lipman, 'Improved tools for biological sequence comparison', *Proc. Natl. Acad. Sci.*, vol. 85, no. 8, pp. 2444–2448, Apr. 1988.
[104] IUPAC-IUB Comm. on Biochem. Nomenclature (CBN), 'Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents', *Biochemistry*, vol. 9, no. 20, pp. 4022–4027, Sep. 1970.
[105] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, 'Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment', *Genome Res.*, vol. 8, no. 3, pp. 175–185, Jan. 1998.
[106] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, Apr. 2010.
[107] H. Li *et al.*, 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.
[108] W. J. Kent *et al.*, 'The Human Genome Browser at UCSC', *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jan. 2002.
[109] P. Danecek *et al.*, 'The variant call format and VCFtools', *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011.
[110] M. G. Grabherr *et al.*, 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, Jul. 2011.
[111] D. N. G. Bruijn, 'A combinatorial problem', *Proc. Sect. Sci. K. Ned. Akad. Van Wet. Te Amst.*, vol. 49, no. 7, pp. 758–764, 1946.
[112] B. J. Haas *et al.*, '*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nat. Protoc.*, vol. 8, no. 8, pp. 1494–1512, Aug. 2013.

[113] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biol.*, vol. 10, no. 3, p. R25, Mar. 2009.
[114] A. Dobin *et al.*, 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.
[115] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, 'Computational methods for transcriptome annotation and quantification using RNA-seq', *Nat. Methods*, vol. 8, no. 6, pp. 469–477, Jun. 2011.
[116] A. R. Quinlan and I. M. Hall, 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.
[117] C. Trapnell *et al.*, 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–515, May 2010.
[118] Y. Xing, A. Resch, and C. Lee, 'The Multiassembly Problem: Reconstructing Multiple Transcript Isoforms From EST Fragment Mixtures', *Genome Res.*, vol. 14, no. 3, pp. 426–441, Jan. 2004.
[119] P. G. Engström *et al.*, 'Systematic evaluation of spliced alignment programs for RNA-seq data', *Nat. Methods*, vol. 10, no. 12, pp. 1185–1191, Dec. 2013.
[120] S. Anders, P. T. Pyl, and W. Huber, 'HTSeq—a Python framework to work with high-throughput sequencing data', *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015.
[121] Y. Liao, G. K. Smyth, and W. Shi, 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014.
[122] J. T. Robinson *et al.*, 'Integrative Genomics Viewer', *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011.
[123] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, 'Analysis and design of RNA sequencing experiments for identifying isoform regulation', *Nat. Methods*, vol. 7, no. 12, pp. 1009–1015, Dec. 2010.
[124] Y. Katz *et al.*, 'Quantitative visualization of alternative exon expression from RNA-seq data', *Bioinformatics*, vol. 31, no. 14, pp. 2400–2402, Jul. 2015.
[125] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008.
[126] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC Bioinformatics*, vol. 11, no. 1, p. 94, Feb. 2010.
[127] J. K. Pickrell *et al.*, 'Understanding mechanisms underlying human gene expression variation with RNA sequencing', *Nature*, vol. 464, no. 7289, pp. 768–772, Apr. 2010.
[128] S. Anders and W. Huber, 'Differential expression analysis for sequence count data', *Genome Biol.*, vol. 11, no. 10, p. R106, Oct. 2010.
[129] M. D. Robinson and A. Oshlack, 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biol.*, vol. 11, no. 3, p. R25, Mar. 2010.
[130] J. P. Venables *et al.*, 'Identification of Alternative Splicing Markers for Breast Cancer', *Cancer Res.*, vol. 68, no. 22, pp. 9525–9531, Nov. 2008.
[131] G. Davy *et al.*, 'Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer', *Eur. J. Hum. Genet.*, vol. 25, no. 10, pp. 1147–1154, Oct. 2017.
[132] S. Tarazona *et al.*, 'Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package', *Nucleic Acids Res.*, vol. 43, no. 21, pp. e140–e140, Dec. 2015.
[133] G. K. Smyth, 'limma: Linear Models for Microarray Data', in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, Eds. New York, NY: Springer New York, 2005, pp. 397–420.
[134] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, 'DEGseq: an R package for identifying differentially expressed genes from RNA-seq data', *Bioinformatics*, vol. 26, no. 1, pp. 136–138, Jan. 2010.
[135] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.

[136] M. I. Love, W. Huber, and S. Anders, ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014.
[137] T. J. Hardcastle and K. A. Kelly, ‘baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data’, *BMC Bioinformatics*, vol. 11, no. 1, p. 422, Aug. 2010.
[138] N. Leng *et al.*, ‘EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments’, *Bioinformatics*, vol. 29, no. 8, pp. 1035–1043, Apr. 2013.
[139] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, ‘Differential analysis of gene regulation at transcript resolution with RNA-seq’, *Nat. Biotechnol.*, vol. 31, no. 1, pp. 46–53, Jan. 2013.
[140] C. Soneson and M. Delorenzi, ‘A comparison of methods for differential expression analysis of RNA-seq data’, *BMC Bioinformatics*, vol. 14, no. 1, p. 91, Mar. 2013.
[141] S. Anders, A. Reyes, and W. Huber, ‘Detecting differential usage of exons from RNA-seq data’, *Genome Res.*, vol. 22, no. 10, pp. 2008–2017, Jan. 2012.
[142] C. Trapnell *et al.*, ‘Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks’, *Nat. Protoc.*, vol. 7, no. 3, pp. 562–578, Mar. 2012.
[143] The 1000 Genomes Project Consortium, ‘A global reference for human genetic variation’, *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015.
[144] M. J. Landrum *et al.*, ‘ClinVar: public archive of interpretations of clinically relevant variants’, *Nucleic Acids Res.*, vol. 44, no. Database issue, pp. D862–D868, Jan. 2016.
[145] M. B. Shapiro and P. Senapathy, ‘RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression.’, *Nucleic Acids Res.*, vol. 15, no. 17, pp. 7155–7174, Sep. 1987.
[146] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud, ‘Human Splicing Finder: an online bioinformatics tool to predict splicing signals’, *Nucleic Acids Res.*, vol. 37, no. 9, pp. e67–e67, May 2009.
[147] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
[148] G. Yeo and C. B. Burge, ‘Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals’, *J. Comput. Biol.*, vol. 11, no. 2–3, pp. 377–394, Mar. 2004.
[149] E. T. Jaynes, ‘Information Theory and Statistical Mechanics’, *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, May 1957.
[150] D. D. Giacomo *et al.*, ‘Functional Analysis of a Large set of BRCA2 exon 7 Variants Highlights the Predictive Value of Hexamer Scores in Detecting Alterations of Exonic Splicing Regulatory Elements’, *Hum. Mutat.*, vol. 34, no. 11, pp. 1547–1557, 2013.
[151] F. Rosenblatt, ‘PRINCIPLES OF NEURODYNAMICS. PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS’, CORNELL AERONAUTICAL LAB INC BUFFALO NY, VG-1196-G-8, Mar. 1961.
[152] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013.
[153] A. Corvelo, M. Hallegger, C. W. J. Smith, and E. Eyras, ‘Genome-Wide Association between Branch Point Properties and Alternative Splicing’, *PLOS Comput. Biol.*, vol. 6, no. 11, p. e1001016, Nov. 2010.
[154] L. Breiman, *Classification and regression trees*. Wadsworth International Group, 1984.
[155] M. Pertea, X. Lin, and S. L. Salzberg, ‘GeneSplicer: a new computational method for splice site prediction’, *Nucleic Acids Res.*, vol. 29, no. 5, pp. 1185–1190, Mar. 2001.
[156] C. Burge and S. Karlin, ‘Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen’, *J. Mol. Biol.*, vol. 268, no. 1, pp. 78–94, Apr. 1997.
[157] L. Breiman, ‘Random Forests’, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[158] L. Breiman, ‘Bagging predictors’, *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[159] P. K. Meher, T. K. Sahu, and A. R. Rao, ‘Prediction of donor splice sites using random forest with a new sequence encoding approach’, *BioData Min.*, vol. 9, no. 1, p. 4, Jan. 2016.
[160] Y. Le Cun, ‘Learning Process in an Asymmetric Threshold Network’, in *Disordered Systems and Biological Organization*, 1986, pp. 233–240.

[161] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ‘Learning Internal Representations by Error Propagation’, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE, ICS-8506, Sep. 1985.
[162] M. G. Reese, F. H. Eeckman, H. Genome, and I. Group, ‘Novel Neural Network Prediction Systems for Human Promoters and Splice Sites’, in *In Gene-Finding and Gene Structure Prediction Workshop*, 1995.
[163] G. E. Hinton, S. Osindero, and Y.-W. Teh, ‘A Fast Learning Algorithm for Deep Belief Nets’, *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, May 2006.
[164] Y. Bengio, ‘Learning Deep Architectures for AI’, *Found Trends Mach Learn*, vol. 2, no. 1, pp. 1–127, Jan. 2009.
[165] K. Xu *et al.*, ‘Show, attend and tell: Neural image caption generation with visual attention’, in *32nd International Conference on Machine Learning, ICML 2015*, 2015, pp. 2048–2057.
[166] B. Signal, B. S. Gloss, M. E. Dinger, T. R. Mercer, and J. Hancock, ‘Machine learning annotation of human branchpoints’, *Bioinformatics*, vol. 34, no. 6, pp. 920–927, Mar. 2018.
[167] J. M. Paggi and G. Bejerano, ‘A sequence-based, deep learning model accurately predicts RNA splicing branchpoints’, *RNA*, p. rna.066290.118, Sep. 2018.
[168] I. Nazari, H. Tayara, and K. T. Chong, ‘Branch Point Selection in RNA Splicing Using Deep Learning’, *IEEE Access*, vol. 7, pp. 1800–1807, 2019.
[169] J. Zuallaert, F. Godin, M. Kim, A. Soete, Y. Saeys, and W. De Neve, ‘SpliceRover: interpretable convolutional neural networks for improved splice site prediction’, *Bioinformatics*, vol. 34, no. 24, pp. 4180–4188, Dec. 2018.
[170] ‘Neural Networks, Manifolds, and Topology -- colah’s blog’. [Online]. Available: https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/. [Accessed: 28-Aug-2019].
[171] P. Divina, A. Kvitkovicova, E. Buratti, and I. Vorechovsky, ‘*Ab initio* prediction of mutation-induced cryptic splice-site activation and exon skipping’, *Eur. J. Hum. Genet.*, vol. 17, no. 6, pp. 759–765, Jun. 2009.
[172] X. Jian, E. Boerwinkle, and X. Liu, ‘In silico prediction of splice-altering single nucleotide variants in the human genome’, *Nucleic Acids Res.*, vol. 42, no. 22, pp. 13534–13544, Dec. 2014.
[173] H. Y. Xiong *et al.*, ‘The human splicing code reveals new insights into the genetic determinants of disease’, *Science*, vol. 347, no. 6218, p. 1254806, Jan. 2015.
[174] I. Vořechovský, ‘Aberrant 3′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization’, *Nucleic Acids Res.*, vol. 34, no. 16, pp. 4630–4641, Sep. 2006.
[175] E. Buratti *et al.*, ‘Aberrant 5′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization’, *Nucleic Acids Res.*, vol. 35, no. 13, pp. 4250–4263, Jul. 2007.
[176] M. H. Zweig and G. Campbell, ‘Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.’, *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, Apr. 1993.
[177] M. Wang and A. Marín, ‘Characterization and prediction of alternative splice sites’, *Gene*, vol. 366, no. 2, pp. 219–227, Feb. 2006.
[178] X. Jian, E. Boerwinkle, and X. Liu, ‘In silico tools for splicing defect prediction: a survey from the viewpoint of end users’, *Genet. Med.*, vol. 16, no. 7, pp. 497–503, Jul. 2014.
[179] X. Liu, X. Jian, and E. Boerwinkle, ‘dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations’, *Hum. Mutat.*, vol. 34, no. 9, pp. E2393–E2402, 2013.
[180] K. Wang, M. Li, and H. Hakonarson, ‘ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data’, *Nucleic Acids Res.*, vol. 38, no. 16, pp. e164–e164, Sep. 2010.
[181] ‘Le cancer du sein - Les cancers les plus fréquents’. [Online]. Available: https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-du-sein. [Accessed: 30-Aug-2019].
[182] ‘INCA - Les cancers en France’. [Online]. Available: http://www.e-cancer.fr/ressources/cancers_en_france/. [Accessed: 30-Aug-2019].
[183] F. Binder-Foucard *et al.*, ‘Cancer incidence and mortality in France over the 1980–2012 period: Solid tumors’, *Epidemiol. Public Health Rev. Epidémiologie Santé Publique*, pp. 95–108, 2014.

[184] H. Kobayashi, S. Ohno, Y. Sasaki, and M. Matsuura, 'Hereditary breast and ovarian cancer susceptibility genes (Review)', *Oncol. Rep.*, vol. 30, no. 3, pp. 1019–1029, Sep. 2013.
[185] Y. Miki *et al.*, 'A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1', *Science*, vol. 266, no. 5182, pp. 66–71, 1994.
[186] R. Wooster *et al.*, 'Identification of the breast cancer susceptibility gene BRCA2', *Nature*, vol. 378, no. 6559, pp. 789–792, Dec. 1995.
[187] K. B. Kuchenbaecker *et al.*, 'Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers', *JAMA*, vol. 317, no. 23, pp. 2402–2416, Jun. 2017.
[188] 'Synthèse - Femmes porteuses d'une mutation de BRCA1 ou BRCA2 / Détection précoce du cancer du sein et des annexes et stratégies de réduction du risque - Ref : RECOBRCASYNTH17'. [Online]. Available: https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Synthese-Femmes-porteuses-d-une-mutation-de-BRCA1-ou-BRCA2-Detection-precoce-du-cancer-du-sein-et-des-annexes-et-strategies-de-reduction-du-risque. [Accessed: 30-Aug-2019].
[189] 'L'oncogénétique en 2017 - consultations et laboratoires - Ref : ADONCOG19'. [Online]. Available: https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/L-oncogenetique-en-2017-consultations-et-laboratoires. [Accessed: 30-Aug-2019].
[190] M. Colombo *et al.*, 'Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium', *Hum. Mol. Genet.*, vol. 23, no. 14, pp. 3666–3680, Jul. 2014.
[191] J. D. Fackenthal *et al.*, 'Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples', *J. Med. Genet.*, vol. 53, no. 8, pp. 548–558, Aug. 2016.
[192] J. Her and S. F. Bunting, 'How cells ensure correct repair of DNA double-strand breaks', *J. Biol. Chem.*, vol. 293, no. 27, pp. 10502–10511, Jun. 2018.
[193] J.-H. Lee and T. T. Paull, 'ATM Activation by DNA Double-Strand Breaks Through the Mre11-Rad50-Nbs1 Complex', 2005.
[194] W. Zhao, C. Wiese, Y. Kwon, R. Hromas, and P. Sung, 'The BRCA Tumor Suppressor Network in Chromosome Damage Repair by Homologous Recombination', *Annu. Rev. Biochem.*, vol. 88, no. 1, pp. 221–245, 2019.
[195] W. Wu, A. Koike, T. Takeshita, and T. Ohta, 'The ubiquitin E3 ligase activity of BRCA1 and its biological functions', *Cell Div.*, vol. 3, no. 1, p. 1, Jan. 2008.
[196] K. I. Savage and D. P. Harkin, 'BRCA1, a "complex" protein involved in the maintenance of genomic stability.', *FEBS J.*, vol. 282, no. 4, pp. 630–646, Feb. 2015.
[197] C. M. Christou and K. Kyriacou, 'BRCA1 and Its Network of Interacting Partners', *Biology*, vol. 2, no. 1, pp. 40–63, Mar. 2013.
[198] F. Zhang *et al.*, 'PALB2 Links BRCA1 and BRCA2 in the DNA-Damage Response', *Curr. Biol.*, vol. 19, no. 6, pp. 524–529, Mar. 2009.
[199] F. Zhang, Q. Fan, K. Ren, and P. R. Andreassen, 'PALB2 Functionally Connects the Breast Cancer Susceptibility Proteins BRCA1 and BRCA2', *Mol. Cancer Res. MCR*, vol. 7, no. 7, pp. 1110–1118, Jul. 2009.
[200] B. Xia *et al.*, 'Control of BRCA2 Cellular and Clinical Functions by a Nuclear Partner, PALB2', *Mol. Cell*, vol. 22, no. 6, pp. 719–729, Jun. 2006.
[201] W. Zhao *et al.*, 'BRCA1-BARD1 promotes RAD51-mediated homologous DNA pairing', *Nature*, vol. 550, no. 7676, pp. 360–365, Oct. 2017.
[202] P. Bork, N. Blomberg, and M. Nilges, 'Internal repeats in the BRCA2 protein sequence', *Nat. Genet.*, vol. 13, no. 1, pp. 22–23, May 1996.
[203] P.-L. Chen, C.-F. Chen, Y. Chen, J. Xiao, Z. D. Sharp, and W.-H. Lee, 'The BRC Repeats in BRCA2 are Critical for RAD51 Binding and Resistance to Methyl Methanesulfonate Treatment', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 9, pp. 5287–5292, 1998.
[204] S. K. Sharan *et al.*, 'Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2.', *Nature*, vol. 386, no. 6627, pp. 804–810, Apr. 1997.
[205] H. Yang *et al.*, 'BRCA2 Function in DNA Binding and Recombination from a BRCA2-DSS1-ssDNA Structure', *Science*, vol. 297, no. 5588, pp. 1837–1848, Sep. 2002.

[206] G. Chatterjee, J. Jimenez-Sainz, T. Presti, T. Nguyen, and R. B. Jensen, 'Distinct binding of BRCA2 BRC repeats to RAD51 generates differential DNA damage sensitivity', *Nucleic Acids Res.*, vol. 44, no. 11, pp. 5256–5270, Jun. 2016.
[207] C. Von Nicolai, Å. Ehlén, C. Martin, X. Zhang, and A. Carreira, 'A second DNA binding site in human BRCA2 promotes homologous recombination.', *12813*, Sep. 2016.
[208] K. Kast *et al.*, 'Prevalence of BRCA1/2 germline mutations in 21 401 families with breast and ovarian cancer', *J. Med. Genet.*, pp. 465–471, 2016.
[209] F. M. Giardiello *et al.*, 'Very high risk of cancer in familial Peutz–Jeghers syndrome', *Gastroenterology*, vol. 119, no. 6, pp. 1447–1453, Dec. 2000.
[210] P. D. P. Pharoah, P. Guilford, and C. Caldas, 'Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families', *Gastroenterology*, vol. 121, no. 6, pp. 1348–1353, Dec. 2001.
[211] T. Walsh *et al.*, 'Spectrum of Mutations in BRCA1, BRCA2, CHEK2, and TP53 in Families at High Risk of Breast Cancer', *Jama J. Am. Med. Assoc.*, vol. 295, no. 12, pp. 1379–1388, Mar. 2006.
[212] V. Bubien *et al.*, 'High cumulative risks of cancer in patients with PTEN hamartoma tumour syndrome', *J. Med. Genet.*, vol. 50, no. 4, pp. 255–263, Apr. 2013.
[213] N. Rahman *et al.*, 'PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene', *Nat. Genet.*, vol. 39, pp. 165–167, 2007.
[214] A. Antoniou *et al.*, 'Breast-Cancer Risk in Families With Mutations in PALB2', *Obstet. Gynecol. Surv.*, vol. 69, no. 11, pp. 659–660, Nov. 2014.
[215] A. Meindl *et al.*, 'Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene', *Nat. Genet.*, vol. 42, no. 5, pp. 410–414, May 2010.
[216] C. Loveday *et al.*, 'Germline mutations in RAD51D confer susceptibility to ovarian cancer.', *PubMed*, Sep. 2011.
[217] S. M. H. Sy, M. S. Y. Huen, and J. Chen, 'PALB2 is an integral component of the BRCA complex required for homologous recombination repair', *Proc. Natl. Acad. Sci.*, vol. 106, no. 17, pp. 7155–7160, Apr. 2009.
[218] J.-Y. Bleuyard, R. Buisson, J.-Y. Masson, and F. Esashi, 'ChAM, a novel motif that mediates PALB2 intrinsic chromatin binding and facilitates DNA repair', *EMBO Rep.*, vol. 13, no. 2, pp. 135–141, Feb. 2012.
[219] T. F. Smith, C. Gaitatzes, K. Saxena, and E. J. Neer, 'The WD repeat: a common architecture for diverse functions', *Trends Biochem. Sci.*, vol. 24, no. 5, pp. 181–185, May 1999.
[220] A. W. Oliver, S. Swift, C. J. Lord, A. Ashworth, and L. H. Pearl, 'Structural basis for recruitment of BRCA2 by PALB2', *EMBO Rep.*, vol. 10, no. 9, pp. 990–996, Sep. 2009.
[221] J. Martino and K. A. Bernstein, 'The Shu complex is a conserved regulator of homologous recombination', *FEMS Yeast Res.*, vol. 16, no. 6, Sep. 2016.
[222] M. R. G. Taylor *et al.*, 'Rad51 Paralogs Remodel Pre-synaptic Rad51 Filaments to Stimulate Homologous Recombination', *Cell*, vol. 162, no. 2, pp. 271–286, Jul. 2015.
[223] S. E. Plon *et al.*, 'Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results', *Hum. Mutat.*, vol. 29, no. 11, pp. 1282–1291, 2008.
[224] E. Mathe, M. Olivier, S. Kato, C. Ishioka, P. Hainaut, and S. V. Tavtigian, 'Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods', *Nucleic Acids Res.*, vol. 34, no. 5, pp. 1317–1325, Mar. 2006.
[225] S. V. Tavtigian *et al.*, 'Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral', *J. Med. Genet.*, vol. 43, no. 4, pp. 295–305, Apr. 2006.
[226] D. F. Easton *et al.*, 'A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer–Predisposition Genes', *Am. J. Hum. Genet.*, vol. 81, no. 5, pp. 873–883, Nov. 2007.
[227] N. M. Lindor *et al.*, 'A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS)', *Hum. Mutat.*, vol. 33, no. 1, pp. 8–21, 2012.

[228] D. Thompson, D. F. Easton, and D. E. Goldgar, 'A Full-Likelihood Method for the Evaluation of Causality of Sequence Variants from Family Data', *Am. J. Hum. Genet.*, vol. 73, no. 3, pp. 652–655, Sep. 2003.
[229] T. Thornton and M. S. McPeek, 'Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test', *Am. J. Hum. Genet.*, vol. 81, no. 2, pp. 321–337, Aug. 2007.
[230] D. G. R. Evans *et al.*, 'A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO', *J. Med. Genet.*, vol. 41, no. 6, pp. 474–480, Jun. 2004.
[231] D. E. Goldgar *et al.*, 'Genetic evidence and integration of various data sources for classifying uncertain variants into a single model', *Hum. Mutat.*, vol. 29, no. 11, pp. 1265–1272, Nov. 2008.
[232] S. Richards *et al.*, 'Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genet. Med. Off. J. Am. Coll. Med. Genet.*, vol. 17, no. 5, pp. 405–424, May 2015.
[233] K. J. Karczewski *et al.*, 'Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes', *bioRxiv*, p. 531210, Jan. 2019.
[234] G. A. Millot, 'A guide for functional analysis of BRCA1 variants of uncertain significance', *Hum. Mutat.*, vol. 33, no. 11, pp. 1526–37, 2012.
[235] L. Guidugli *et al.*, 'Functional Assays for Analysis of Variants of Uncertain Significance in BRCA2', *Hum. Mutat.*, vol. 35, no. 2, pp. 151–164, Feb. 2014.
[236] S. G. Kuznetsov, P. Liu, and S. K. Sharan, 'Mouse ES-cell-based functional assay to evaluate mutations in BRCA2', *Nat. Med.*, vol. 14, no. 8, pp. 875–881, Aug. 2008.
[237] S. Chang, K. Biswas, B. Martin, S. Stauffer, and S. Sharan, 'Expression of human BRCA1 variants in mouse ES cells allows functional analysis of BRCA1 mutations', *J. Clin. Invest.*, vol. 119, no. 10, pp. 3160–3171, Oct. 2009.
[238] P. Bouwman *et al.*, 'A High-Throughput Functional Complementation Assay for Classification of BRCA1 Missense Variants'.
[239] D. J. R. Ransburgh, N. Chiba, C. Ishioka, A. E. Toland, and J. D. Parvin, 'Identification of Breast Tumor Mutations in BRCA1 That Abolish Its Function in Homologous DNA Recombination', *Cancer Res.*, vol. 70, no. 3, pp. 988–995, Feb. 2010.
[240] L. Guidugli *et al.*, 'A Classification Model for BRCA2 DNA Binding Domain Missense Variants Based on Homology-Directed Repair Activity', *Cancer Res.*, vol. 73, no. 1, pp. 265–275, Jan. 2013.
[241] D. J. Farrugia *et al.*, 'Functional Assays for Classification of BRCA2 Variants of Uncertain Significance', *Cancer Res.*, vol. 68, no. 9, pp. 3523–3531, May 2008.
[242] Z. Kais, N. Chiba, C. Ishioka, and J. D. Parvin, 'Functional differences among BRCA1 missense mutations in the control of centrosome duplication', *Oncogene*, vol. 31, no. 6, pp. 799–804, Feb. 2012.
[243] C. Béroud *et al.*, 'BRCA Share: A Collection of Clinical BRCA Gene Variants', *Hum. Mutat.*, vol. 37, no. 12, pp. 1318–1328, 2016.
[244] A. B. Spurdle *et al.*, 'ENIGMA—Evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes', *Hum. Mutat.*, vol. 33, no. 1, pp. 2–7, 2012.
[245] M. T. Parsons *et al.*, 'Large scale multifactorial likelihood quantitative analysis of BRCA1 and BRCA2 variants: An ENIGMA resource to support clinical variant classification', *Hum. Mutat.*, vol. 0, no. ja.
[246] M. S. Cline *et al.*, 'BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2', *PLOS Genet.*, vol. 14, no. 12, p. e1007752, Dec. 2018.
[247] J. Balmaña *et al.*, 'Conflicting Interpretation of Genetic Variants and Cancer Risk by Commercial Laboratories as Assessed by the Prospective Registry of Multiplex Testing', *J. Clin. Oncol.*, vol. 34, no. 34, pp. 4071–4078, Sep. 2016.

[248] S. M. Harrison *et al.*, 'Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar', *Genet. Med.*, vol. 19, no. 10, pp. 1096–1104, Oct. 2017.
[249] M. Colombo *et al.*, 'Comparative In Vitro and In Silico Analyses of Variants in Splicing Regions of BRCA1 and BRCA2 Genes and Characterization of Novel Pathogenic Mutations', *PLOS ONE*, vol. 8, no. 2, p. e57173, Feb. 2013.
[250] B. Wappenschmidt *et al.*, 'Analysis of 30 Putative BRCA1 Splicing Mutations in Hereditary Breast and Ovarian Cancer Families Identifies Exonic Splice Site Mutations That Escape In Silico Prediction', *PLOS ONE*, vol. 7, no. 12, p. e50800, Dec. 2012.
[251] S. M. Caputo *et al.*, 'Full in-frame exon 3 skipping of BRCA2 confers high risk of breast and/or ovarian cancer', *Oncotarget*, vol. 9, no. 25, pp. 17334–17348, Apr. 2018.
[252] E. Fraile-Bethencourt *et al.*, 'Mis-splicing in breast cancer: identification of pathogenic BRCA2 variants by systematic minigene assays', *J. Pathol.*, vol. 248, no. 4, pp. 409–420, 2019.
[253] L. Li *et al.*, 'Functional Redundancy of Exon 12 of BRCA2 Revealed by a Comprehensive Analysis of the c.6853A>G (p.I2285V) Variant', *Hum. Mutat.*, vol. 30, no. 11, pp. 1543–1550, Nov. 2009.
[254] C. Houdayer *et al.*, 'Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants', *Hum. Mutat.*, vol. 33, no. 8, pp. 1228–1238, Aug. 2012.
[255] M. de la Hoya *et al.*, 'Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms', *Hum. Mol. Genet.*, vol. 25, no. 11, pp. 2256–2268, Jun. 2016.
[256] A. Goyenvalle *et al.*, 'Rescue of Dystrophic Muscle Through U7 snRNA-Mediated Exon Skipping', *Science*, vol. 306, no. 5702, pp. 1796–1799, Dec. 2004.
[257] L. Meulemans, 'Développement de nouveaux essais fonctionnels pour l'analyse de l'impact des variations dans des gènes de prédisposition aux cancers.'
[258] 'Lawsuit raises questions about variant interpretation and communication', *Am. J. Med. Genet. A.*, vol. 173, no. 4, pp. 838–839, Apr. 2017.
[259] C. Houdayer *et al.*, 'Evaluation of in silico splice tools for decision-making in molecular diagnosis', *Hum. Mutat.*, vol. 29, no. 7, pp. 975–982, 2008.
[260] R. Leman *et al.*, 'Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort', *Nucleic Acids Res.*, vol. 46, no. 21, pp. 11656–11657, Nov. 2018.
[261] Q. Zhang *et al.*, 'BPP: a sequence-based algorithm for branch point prediction', *Bioinformatics*, vol. 33, no. 20, pp. 3166–3172, Oct. 2017.
[262] R. Leman *et al.*, 'Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants', *BMC Genomics*, In revision.
[263] J. Vaquero-Garcia *et al.*, 'A new view of transcriptome complexity and regulation through the lens of local splicing variations.', *eLife*, vol. 5, pp. e11752–e11752, 2016.
[264] I. Lopez-Perolio *et al.*, 'Alternative splicing and ACMG-AMP-2015-based classification of PALB2 genetic variants: an ENIGMA report', *J. Med. Genet.*, p. jmedgenet-2018-105834, Mar. 2019.
[265] S. Caputo, L. Benboudjema, O. Sinilnikova, E. Rouleau, C. Béroud, and R. Lidereau, 'Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases', *Nucleic Acids Res.*, vol. 40, no. D1, pp. D992–D1002, Jan. 2012.
[266] C. Szabo, A. Masiello, J. F. Ryan, and L. C. Brody, 'The Breast Cancer Information Core: Database design, structure, and scope', *Hum. Mutat.*, vol. 16, no. 2, pp. 123–131, Aug. 2000.
[267] D. Baralle, A. Lucassen, and E. Buratti, 'Missed threads: The impact of pre-mRNA splicing defects on clinical practice', *EMBO Rep.*, vol. 10, no. 8, pp. 810–816, Aug. 2009.
[268] M. P. Vallée *et al.*, 'Adding In Silico Assessment of Potential Splice Aberration to the Integrated Evaluation of BRCA Gene Unclassified Variants', *Hum. Mutat.*, vol. 37, no. 7, pp. 627–639, Jul. 2016.
[269] L. C. Walker *et al.*, 'Evaluation of a 5-Tier Scheme Proposed for Classification of Sequence Variants Using Bioinformatic and Splicing Assay Data: Inter-Reviewer Variability and

Promotion of Minimum Reporting Guidelines', *Hum. Mutat.*, vol. 34, no. 10, pp. 1424–1431, Oct. 2013.
[270] D. J. Sanz *et al.*, 'A High Proportion of DNA Variants of BRCA1 and BRCA2 Is Associated with Aberrant Splicing in Breast/Ovarian Cancer Patients', *Clin. Cancer Res.*, vol. 16, no. 6, pp. 1957–1967, Mar. 2010.
[271] I. Callebaut *et al.*, 'Comprehensive functional annotation of 18 missense mutations found in suspected hemochromatosis type 4 patients', *Hum. Mol. Genet.*, vol. 23, no. 17, pp. 4479–4490, Sep. 2014.
[272] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, 'ROCR: visualizing classifier performance in R', *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, Oct. 2005.
[273] M. Burset, I. A. Seledtsov, and V. V. Solovyev, 'SpliceDB: database of canonical and non-canonical mammalian splice sites', *Nucleic Acids Res.*, vol. 29, no. 1, p. 255, Jan. 2001.
[274] S. Ke *et al.*, 'Saturation mutagenesis reveals manifold determinants of exon definition', *Genome Res.*, vol. 28, no. 1, pp. 11–24, Jan. 2018.
[275] Y. Lee and D. C. Rio, 'Mechanisms and Regulation of Alternative Pre-mRNA Splicing', *Annu. Rev. Biochem.*, vol. 84, no. 1, pp. 291–323, 2015.
[276] O. Soukarieh *et al.*, 'Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools', *PLOS Genet.*, vol. 12, no. 1, p. e1005756, Jan. 2016.
[277] P. Julien, B. Miñana, P. Baeza-Centurion, J. Valcárcel, and B. Lehner, 'The complete local genotype–phenotype landscape for the alternative splicing of a human exon', *Nat. Commun.*, vol. 7, p. 11558, May 2016.
[278] L. A. Chasin, 'Searching for splicing motifs', *Adv. Exp. Med. Biol.*, vol. 623, pp. 85–106, 2007.
[279] Ø. L. Holla *et al.*, 'Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: Comparison of wet-lab and bioinformatics analyses', *Mol. Genet. Metab.*, vol. 96, no. 4, pp. 245–252, Apr. 2009.
[280] J. C. Théry *et al.*, 'Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the *BRCA* genes', *Eur. J. Hum. Genet.*, vol. 19, no. 10, pp. 1052–1058, Oct. 2011.
[281] M. P. G. Vreeswijk *et al.*, 'Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs', *Hum. Mutat.*, vol. 30, no. 1, pp. 107–114, Jan. 2009.
[282] P. J. Whiley *et al.*, 'Splicing and multifactorial analysis of intronic BRCA1 and BRCA2 sequence variants identifies clinically significant splicing aberrations up to 12 nucleotides from the intron/exon boundary', *Hum. Mutat.*, vol. 32, no. 6, pp. 678–687, 2011.
[283] R. Tang, D. O. Prosser, and D. R. Love, 'Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions', *Advances in Bioinformatics*, 2016. [Online]. Available: https://www.hindawi.com/journals/abi/2016/5614058/. [Accessed: 17-Jan-2018].
[284] M. S. Jurica and M. J. Moore, 'Pre-mRNA Splicing: Awash in a Sea of Proteins', *Mol. Cell*, vol. 12, no. 1, pp. 5–14, Jul. 2003.
[285] K. Gao, A. Masuda, T. Matsuura, and K. Ohno, 'Human branch point consensus sequence is yUnAy', *Nucleic Acids Res.*, vol. 36, no. 7, pp. 2257–2267, Apr. 2008.
[286] C. L. Will and R. Lührmann, 'Spliceosome Structure and Function', *Cold Spring Harb. Perspect. Biol.*, vol. 3, no. 7, p. a003707, Jan. 2011.
[287] L. D. Conti, M. Baralle, and E. Buratti, 'Exon and intron definition in pre-mRNA splicing', *Wiley Interdiscip. Rev. RNA*, vol. 4, no. 1, pp. 49–60, 2013.
[288] R. Castelo and R. Guigó, 'Splice site identification by idlBNs', *Bioinformatics*, vol. 20, no. suppl_1, pp. i69–i76, Aug. 2004.
[289] C. Gooding, F. Clark, M. C. Wollerton, S.-N. Grellscheid, H. Groom, and C. W. Smith, 'A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones', *Genome Biol.*, vol. 7, no. 1, p. R1, Jan. 2006.
[290] J. M. B. Pineda and R. K. Bradley, 'Most human introns are recognized via multiple and tissue-specific branchpoints', *Genes Dev.*, Apr. 2018.

[291] M. Briese *et al.*, 'Transcriptome-wide profiling of mammalian spliceosome and branchpoints with iCLIP', *bioRxiv*, p. 353599, Jun. 2018.
[292] C. W. Smith, T. T. Chu, and B. Nadal-Ginard, 'Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns.', *Mol. Cell. Biol.*, vol. 13, no. 8, pp. 4939–4952, Aug. 1993.
[293] J. Cheng *et al.*, 'MMSplice: modular modeling improves the predictions of genetic variant effects on splicing', *Genome Biol.*, vol. 20, no. 1, p. 48, Mar. 2019.
[294] J. Královičová, H. Lei, and I. Vořechovský, 'Phenotypic consequences of branch point substitutions', *Hum. Mutat.*, vol. 27, no. 8, pp. 803–813, 2006.
[295] D. R. Zerbino *et al.*, 'Ensembl 2018', *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, Jan. 2018.
[296] J. Wen, J. Wang, Q. Zhang, and D. Guo, 'A heuristic model for computational prediction of human branch point sequence', *BMC Bioinformatics*, vol. 18, no. 1, p. 459, Oct. 2017.
[297] L. Grodecká, E. Buratti, and T. Freiberger, 'Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics?', *Int. J. Mol. Sci.*, vol. 18, no. 8, p. 1668, Aug. 2017.
[298] K. Jaganathan *et al.*, 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, vol. 176, no. 3, pp. 535-548.e24, Jan. 2019.
[299] S. N. Teraoka *et al.*, 'Splicing Defects in the Ataxia-Telangiectasia Gene, ATM: Underlying Mutations and Consequences', *Am. J. Hum. Genet.*, vol. 64, no. 6, pp. 1617–1631, Jun. 1999.
[300] I. Tournier *et al.*, 'A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects', *Hum. Mutat.*, vol. 29, no. 12, pp. 1412–1424, 2008.
[301] W. F. Mueller, L. S. Z. Larsen, A. Garibaldi, G. W. Hatfield, and K. J. Hertel, 'The Silent Sway of Splicing by Synonymous Substitutions', *J. Biol. Chem.*, vol. 290, no. 46, pp. 27700–27711, Nov. 2015.
[302] T. V. O. Hansen, A. Y. Steffensen, L. Jønson, M. K. Andersen, B. Ejlertsen, and F. C. Nielsen, 'The silent mutation nucleotide 744 G → A, Lys172Lys, in exon 6 of BRCA2 results in exon skipping', *Breast Cancer Res. Treat.*, vol. 119, no. 3, pp. 547–550, Feb. 2010.
[303] R. D. Brandão, K. van Roozendaal, D. Tserpelis, E. G. García, and M. J. Blok, 'Characterisation of unclassified variants in the BRCA1/2 genes with a putative effect on splicing', *Breast Cancer Res. Treat.*, vol. 129, no. 3, pp. 971–982, Oct. 2011.
[304] A. Acedo, C. Hernández-Moro, Á. Curiel-García, B. Díez-Gómez, and E. A. Velasco, 'Functional Classification of BRCA2 DNA Variants by Splicing Assays in a Large Minigene with 9 Exons', *Hum. Mutat.*, vol. 36, no. 2, pp. 210–221, 2015.
[305] C. Zhang, B. Zhang, L.-L. Lin, and S. Zhao, 'Evaluation and comparison of computational tools for RNA-seq isoform quantification', *BMC Genomics*, vol. 18, no. 1, p. 583, Aug. 2017.
[306] R. Chaligné *et al.*, 'The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer', *Genome Res.*, vol. 25, no. 4, pp. 488–503, Jan. 2015.
[307] D. F. Easton *et al.*, 'Gene-panel sequencing and the prediction of breast-cancer risk', *N. Engl. J. Med.*, vol. 372, no. 23, pp. 2243–2257, Jun. 2015.
[308] A. Desmond *et al.*, 'Clinical Actionability of Multigene Panel Testing for Hereditary Breast and Ovarian Cancer Risk Assessment', *JAMA Oncol.*, vol. 1, no. 7, pp. 943–951, Oct. 2015.
[309] R. Graffeo, L. Livraghi, O. Pagani, A. Goldhirsch, A. H. Partridge, and J. E. Garber, 'Time to incorporate germline multigene panel testing into breast and ovarian cancer patient care', *Breast Cancer Res. Treat.*, vol. 160, no. 3, pp. 393–410, Dec. 2016.
[310] M. B. Daly *et al.*, 'NCCN Guidelines Insights: Genetic/Familial High-Risk Assessment: Breast and Ovarian, Version 2.2017', *J. Natl. Compr. Cancer Netw. JNCCN*, vol. 15, no. 1, pp. 9–20, Jan. 2017.
[311] D. M. Eccles *et al.*, 'BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance', *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO*, vol. 26, no. 10, pp. 2057–2065, Oct. 2015.
[312] V. Dosil *et al.*, 'Alternative splicing and molecular characterization of splice site variants: BRCA1 c.591C>T as a case study', *Clin. Chem.*, vol. 56, no. 1, pp. 53–61, Jan. 2010.

[313] J. M. Mudge *et al.*, 'The origins, evolution, and functional potential of alternative splicing in vertebrates', *Mol. Biol. Evol.*, vol. 28, no. 10, pp. 2949–2959, Oct. 2011.
[314] P. J. Byrd *et al.*, 'A Hypomorphic PALB2 Allele Gives Rise to an Unusual Form of FA-N Associated with Lymphoid Tumour Development', *PLoS Genet.*, vol. 12, no. 3, p. e1005945, Mar. 2016.
[315] T. A. Thanaraj and F. Clark, 'Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions', *Nucleic Acids Res.*, vol. 29, no. 12, pp. 2581–2593, Jun. 2001.
[316] E. Kim, A. Magen, and G. Ast, 'Different levels of alternative splicing among eukaryotes', *Nucleic Acids Res.*, vol. 35, no. 1, pp. 125–131, 2007.
[317] S. Djebali *et al.*, 'Landscape of transcription in human cells', *Nature*, vol. 489, no. 7414, pp. 101–108, Sep. 2012.
[318] B. J. Blencowe, 'Alternative splicing: new insights from global analyses', *Cell*, vol. 126, no. 1, pp. 37–47, Jul. 2006.
[319] M. L. Tress, F. Abascal, and A. Valencia, 'Most Alternative Isoforms Are Not Functionally Important', *Trends Biochem. Sci.*, vol. 42, no. 6, pp. 408–410, 2017.
[320] B. J. Blencowe, 'The Relationship between Alternative Splicing and Proteomic Complexity', *Trends Biochem. Sci.*, vol. 42, no. 6, pp. 407–408, 2017.
[321] K. J. Niklas, S. E. Bondos, A. K. Dunker, and S. A. Newman, 'Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications', *Front. Cell Dev. Biol.*, vol. 3, p. 8, 2015.
[322] P. J. Whiley *et al.*, 'Comparison of mRNA Splicing Assay Protocols across Multiple Laboratories: Recommendations for Best Practice in Standardized Clinical Testing', *Clin. Chem.*, vol. 60, no. 2, pp. 341–352, Feb. 2014.
[323] A. A. Tesoriero *et al.*, 'Molecular characterization and cancer risk associated with BRCA1 and BRCA2 splice site variants identified in multiple-case breast cancer families', *Hum. Mutat.*, vol. 26, no. 5, p. 495, Nov. 2005.
[324] M. H. Nieuwenhuis and H. F. A. Vasen, 'Correlations between mutation site in APC and phenotype of familial adenomatous polyposis (FAP): a review of the literature', *Crit. Rev. Oncol. Hematol.*, vol. 61, no. 2, pp. 153–161, Feb. 2007.
[325] E. T. Rosenthal *et al.*, 'Exceptions to the rule: case studies in the prediction of pathogenicity for genetic variants in hereditary cancer genes', *Clin. Genet.*, vol. 88, no. 6, pp. 533–541, Dec. 2015.
[326] J.-Y. Park, F. Zhang, and P. R. Andreassen, 'PALB2: the hub of a network of tumor suppressors involved in DNA damage responses', *Biochim. Biophys. Acta*, vol. 1846, no. 1, pp. 263–275, Aug. 2014.
[327] S. Erkelenz, S. Theiss, M. Otte, M. Widera, J. O. Peter, and H. Schaal, 'Genomic HEXploring allows landscaping of novel potential splicing regulatory elements', *Nucleic Acids Res.*, vol. 42, no. 16, pp. 10681–10697, Sep. 2014.
[328] H. Tubeuf *et al.*, 'Recommendations to prioritize genetic variants for RNA analyses derived from a large-scale performance evaluation of splicing regulation-predictors', In revision.
[329] L. Crotti *et al.*, 'A KCNH2 branch point mutation causing aberrant splicing contributes to an explanation of genotype-negative long QT syndrome', *Heart Rhythm*, vol. 6, no. 2, pp. 212–218, Feb. 2009.
[330] J. Lonsdale *et al.*, 'The Genotype-Tissue Expression (GTEx) project', *Nat. Genet.*, vol. 45, pp. 580–585, May 2013.
[331] R. Leman *et al.*, 'SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data', *Bioinformatics*, Accepted.
[332] S. Anders *et al.*, 'Count-based differential expression analysis of RNA sequencing data using R and Bioconductor', *Nat. Protoc.*, vol. 8, no. 9, pp. 1765–1786, Aug. 2013.
[333] R. D. Brandão *et al.*, 'Targeted RNA-seq successfully identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes', *Int. J. Cancer*, vol. 145, no. 2, pp. 401–414, 2019.
[334] R. H. Osborne, J. L. Hopper, J. A. Kirk, G. Chenevix-Trench, H. J. Thorne, and J. F. Sambrook, 'kConFab: a research resource of Australasian breast cancer families', *Med. J. Aust.*, vol. 172, no. 9, pp. 463–464, 2000.

[335] S. Farber-Katz *et al.*, 'Quantitative Analysis of BRCA1 and BRCA2 Germline Splicing Variants Using a Novel RNA-Massively Parallel Sequencing Assay', *Front. Oncol.*, vol. 8, 2018.
[336] J. Soukupova *et al.*, 'Validation of CZECANCA (CZEch CAncer paNel for Clinical Application) for targeted NGS-based analysis of hereditary cancer syndromes.', *PloS One*, vol. 13, no. 4, pp. e0195761–e0195761, 2018.
[337] A. Rohlin *et al.*, 'Expanding the genotype–phenotype spectrum in hereditary colorectal cancer by gene panel testing', *Fam. Cancer*, vol. 16, no. 2, pp. 195–203, Apr. 2017.
[338] S. Kujawa *et al.*, 'A Method for the Identification of Variants in Alzheimer's Disease Candidate Genes and Transcripts Using Hybridization Capture Combined with Long-Read Sequencing', 2016.
[339] M. Tardaguila *et al.*, 'SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification', *Genome Res.*, vol. 28, no. 3, pp. 396–411, Jan. 2018.
[340] L. Duran-Lozano *et al.*, 'Alternative transcript imbalance underlying breast cancer susceptibility in a family carrying PALB2 c.3201+5G>T', *Breast Cancer Res. Treat.*, vol. 174, no. 2, pp. 543–550, Apr. 2019.
[341] L. Castéra *et al.*, 'Landscape of pathogenic variations in a panel of 34 genes and cancer risk estimation from 5131 HBOC families', *Genet. Med.*, vol. 20, no. 12, pp. 1677–1686, Dec. 2018.
[342] A. M. Gruver, B. D. Yard, C. McInnes, C. Rajesh, and D. L. Pittman, 'Functional characterization and identification of mouse Rad51d splice variants', *BMC Mol. Biol.*, vol. 10, no. 1, p. 27, Mar. 2009.
[343] R. A. Baldock *et al.*, 'RAD51D splice variants and cancer-associated mutations reveal XRCC2 interaction to be critical for homologous recombination', *DNA Repair*, vol. 76, pp. 99–107, Apr. 2019.
[344] T. Ludwig, D. L. Chapman, V. E. Papaioannou, and A. Efstratiadis, 'Targeted mutations of breast cancer susceptibility gene homologs in mice: lethal phenotypes of Brca1, Brca2, Brca1/Brca2, Brca1/p53, and Brca2/p53 nullizygous embryos.', *Genes Dev.*, vol. 11, no. 10, pp. 1226–1241, May 1997.
[345] X. Xu *et al.*, 'Genetic interactions between tumor suppressors Brca1 and p53 in apoptosis, cell cycle and tumorigenesis', *Nat. Genet.*, vol. 28, no. 3, pp. 266–271, Jul. 2001.
[346] Y. Wang *et al.*, 'The BRCA1-Δ11q Alternative Splice Isoform Bypasses Germline Mutations and Promotes Therapeutic Resistance to PARP Inhibition and Cisplatin', *Cancer Res.*, vol. 76, no. 9, pp. 2778–2790, May 2016.
[347] B. Wappenschmidt *et al.*, 'Analysis of 30 Putative BRCA1 Splicing Mutations in Hereditary Breast and Ovarian Cancer Families Identifies Exonic Splice Site Mutations That Escape In Silico Prediction', *PLOS ONE*, vol. 7, no. 12, p. e50800, Dec. 2012.
[348] A. Byrjalsen, A. Y. Steffensen, T. v O. Hansen, K. Wadt, and A.-M. Gerdes, 'Classification of the spliceogenic BRCA1 c.4096+3A>G variant as likely benign based on cosegregation data and identification of a healthy homozygous carrier', *Clin. Case Rep.*, vol. 5, no. 6, pp. 876–879, 2017.
[349] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz, 'Alternative splicing: a pivotal step between eukaryotic transcription and translation', *Nat. Rev. Mol. Cell Biol.*, vol. 14, no. 3, pp. 153–165, Mar. 2013.
[350] 'Dr. Mary-Claire King Proposes Population Screening in All Young Women for BRCA Mutations - The ASCO Post'. [Online]. Available: https://www.ascopost.com/issues/february-10-2015/dr-mary-claire-king-proposes-population-screening-in-all-young-women-for-brca-mutations/. [Accessed: 23-Sep-2019].
[351] P. D. Beitsch *et al.*, 'Underdiagnosis of Hereditary Breast Cancer: Are Genetic Testing Guidelines a Tool or an Obstacle?', *J. Clin. Oncol.*, vol. 37, no. 6, pp. 453–460, Dec. 2018.

LIENS DE VULGARISATION SCIENTIFIQUE :

Deep-learning : https://www.youtube.com/watch?v=trWrEWfhTVg

p-value : https://www.youtube.com/watch?v=xVIt51ybvu0

Variable aléatoire : https://www.youtube.com/watch?v=ntNF_VIYexQ

ANNEXES

I. ANNEXE A SUPPLEMENTARY INFORMATION: Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort.

1. Supplementary methods

Model stability was measured by leave-one-out method. Here we used bootstrap approach (Table S8 and Figure S1). Adding to this, we processed to cross-validation with all variants in *BRCA1* and *BRCA2* genes (n = 305). We realized 10,000 iterations with random distribution of 205 variants as training set and 100 variants as validation set. We added 90 variants from other genes to this validation set, a total of 190 variants in validation set. For each iteration, we compared model with only MES or SSF-like to model with MES and SSF-like (Table S9), with comparison of discriminant capacity of SPiCE model to SSF-like and MES alone by ROC analysis (Figure S2).

From validation dataset, reliability diagram was used (Figure S3); proportion of variants with splice effect was represented according to probability fitted by our model. We processed also to similar analysis to reliability diagram but we compared proportion of variants with splice effect for each nucleotidic position of splice site.

Akaike Information Criterion (AIC) takes into account the model likelihood and weights it by the number of k parameters ($AIC = 2 \times (k + 1) - 2 \times \ln(likelihood)$). Thus, AIC provided us with the best model with a minimum of parameters. The Bayesian Information Criterion (BIC) differs from AIC by taking into account the number of n observations ($BIC = 2 \times \ln(n) \times (k + 1) - 2 \times \ln(likelihood)$). This supplementary adjustment avoids a bias linked to the subset size of variables. As a result, BIC is more adapted to a descriptive approach than a predictive approach for which AIC is more relevant. Anyway, it can be observed that AIC and BIC give the same results in the majority of cases. Our model illustrates this observation as shown in table S7, where variations of AIC and BIC were similar.

2. Supplementary tables and figures

Table S1, table S2 and table S3 are in excel format available at https://academic.oup.com/nar/article/46/21/11656/5128933#supplementary-data.

Table S4: Splicing effect observed for variants in our datasets (n = 395)

		Splicing alteration			
			5'/3' alternative splice site		
Datasets	No alteration	Exon skipping	5'	3'	Multiple alteration
Training set	47 (33.1)	61 (43.0)	13 (9.1)	15 (10.6)	6 (4.2)
BRCA1/BRCA2 validation set	28 (17.2)	77 (47.2)	23 (14.1)	14 (8.6)	21 (12.9)
Non *BRCA* validation set	17 (18.9)	38 (42.2)	7 (7.8)	6 (6.7)	22 (24.4)
Total	92 (23.3)	176 (44.6)	43 (10.9)	35 (8.9)	49 (12.4)

Table S5: Correlation coefficients between different scores

	SSF-like	MES	HSF	GS	NNS
SSF-like	1	0.71	0.80	0.47	0.60
MES		1	0.77	0.48	0.87
HSF			1	0.43	0.73
GS				1	0.41
NNS					1

SSF-like: SpliceSite Finder-like, MES: MaxEntScan, HSF: Human Splicing Finder, GS: GeneSplicer, NNS: Neural Network Splice

Table S6: Splicing predictions and effects for 51 *BRCA1/2* variants collected by UGG group since 2012 with guidelines of Houdayer et coll

	With splicing alteration	Without splicing alteration
ΔSSF > 5% and ΔMES > 15%	26	0
ΔSSF < 5% or ΔMES < 15%	9	16

ΔSSF: relative variation score of SpliceSite Finder-like, ΔMES: relative variation score of MaxEntScan

Table S7: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for variables and combination of variables with p-value of Likelihood Ratio Test (LRT)

Variables	AIC	BIC	p-value of LRT
MES	63.46	69.37	< 2.2e-16
SSF-like	81.39	87.61	< 2.2e-16
HSF	94.91	100.82	< 2.2e-16
NNS	108.14	114.05	< 2.2e-16
Invariant site	140.99	146.9	4.663e-11
GS	146.67	152.59	8.554e-10
Splice site	176.34	182.25	0.00477
Gene	184.30	190.21	0.9519
MES + SSF-like	49.50	58.41	6.61e-05
MES + HSF	63.81	72.68	0.1994
MES + NNS	65.37	74.23	0.7591
MES + Invariant Site	65.33	74.20	0.7169
MES + GS	63.42	72.29	0.1536
MES + Splice Site	64.81	73.67	0.4182
MES + Gene	64.43	73.30	0.3105

SSF-like: SpliceSite Finder-like, MES: MaxEntScan, HSF: Human Splicing Finder, GS: GeneSplicer, NNS: Neural Network Splice, Invariant site: 1 for variant in AG/GU sequence and 0 for variant outside, Splice site: 0 for variant in acceptor splice site and 1 for variant in donor splice site, Gene: 0 for variant in *BRCA1* and 1 for variant in *BRCA2*.

Table S8: Results of bootstrap validation

Parameters	Value	CV (%)	$IC_{95\%}$
β_0	-3.59	1.82	[-3.89; -3.52]
β_{MES}	-8.21	1.79	[-8.80; -8.08]
$\beta_{SSF\text{-}like}$	-32.32	3.49	[-35.85; -31.27]
Th_{Se}	0.115	4.19	[0.100; 0.121]
Th_{Sp}	0.747	1.71	[0.721; 0.754]

Parameters are considered as stable if CV < 5%
CV: Coefficient Variation; $IC_{95\%}$: Interval of Confidence at 95%, β_0: intercept, β_{MES}: parameter of MES score, $\beta_{SSF\text{-}like}$: parameter of SSF-like score, Th_{Se}: optimal sensitivity decision threshold, Th_{Se}: optimal specificity decision threshold

Table S9: Results of cross-validation (10,000 iterations)

	MES only model	MES + SSF-like model (SPiCE)
AIC, average[min, max]†	85.73 [46.11; 108.91]	71.07 [26.48; 91.90]
AUC of ROC curve, average[min, max]	0.977 [0.944; 0.996]	0.983 [0.957; 0.997]
Accuracy, average[min, max]‡	0.939 [0.9; 0.979]	0.955 [0.916; 0.984]

†p-value of likelihood ratio test was lower to 5% in more 99% of iterations
‡Maximum accuracy from each ROC analysis
AIC: Akaike Index Criterion, AUC: Area Under the Curve, MES: MaxEntScan, ROC: Receiver Operating Characteristics, SSF-like: Splice Site Finder-like

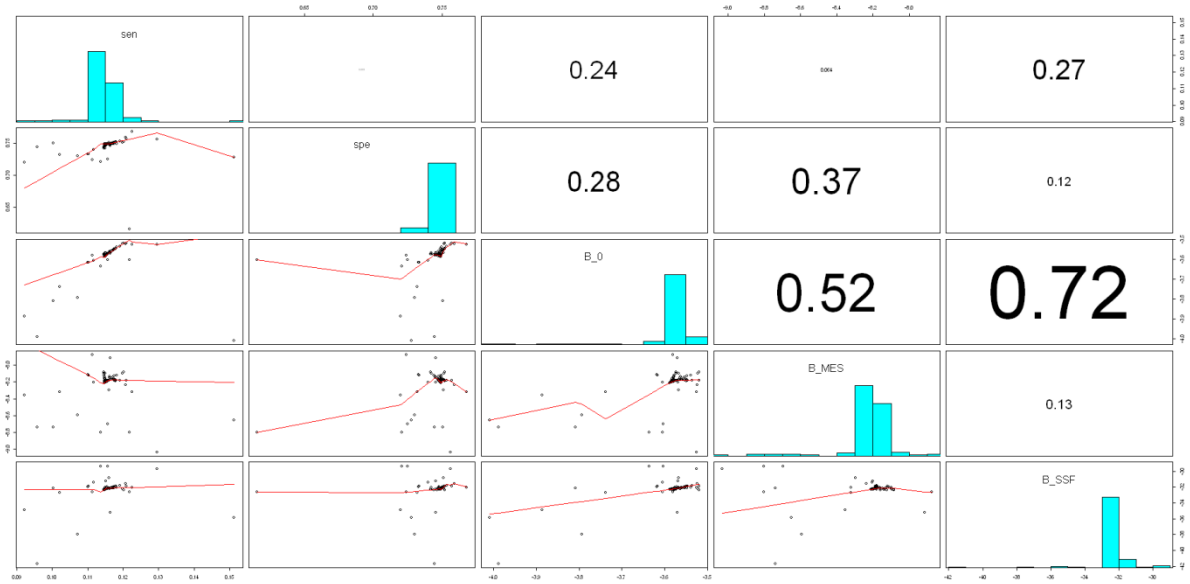


Figure S1: Results of bootstrap method. Pearson correlation coefficients between model parameters are shown in the upper right corner. In the lower left corner, the plot of results between two models parameters are shown with tendency in red, in diagonally, histogram of variables. Bootstrap validation also revealed the independence between thresholds and between model parameters excepted for β_0 and β_{SSF} (correlation coefficient: 0.72).
Sen: threshold of optimal sensitivity, Spe: threshold of optimal specificity, B_0: intercept, B_MES: parameter of MES score, B_SSF: parameter of SSF score

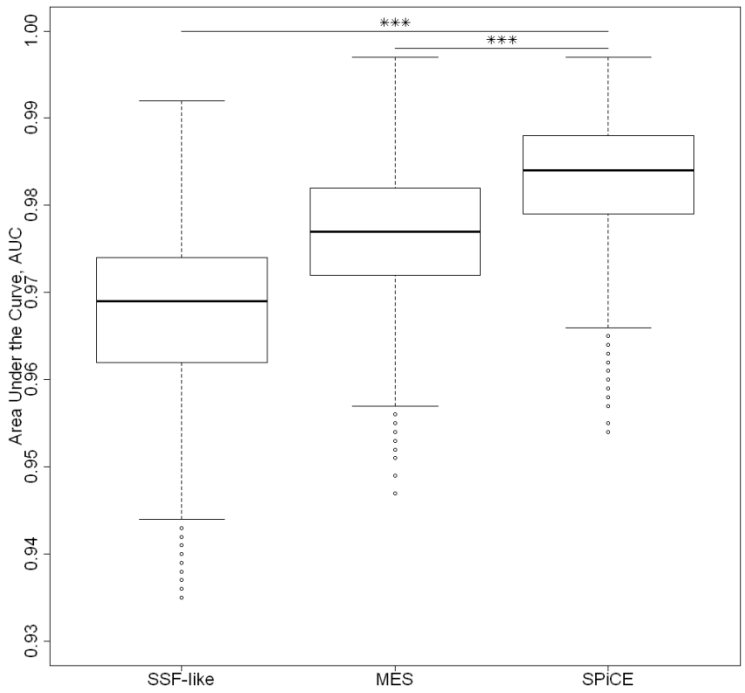


Figure S2: Area Under the Curve (AUC) of ROC curve obtained from cross-validation of SPiCE model from validations sets (10,000 iters). ***: p-value < 2.2e-16 of student test (unilateral and paired test)

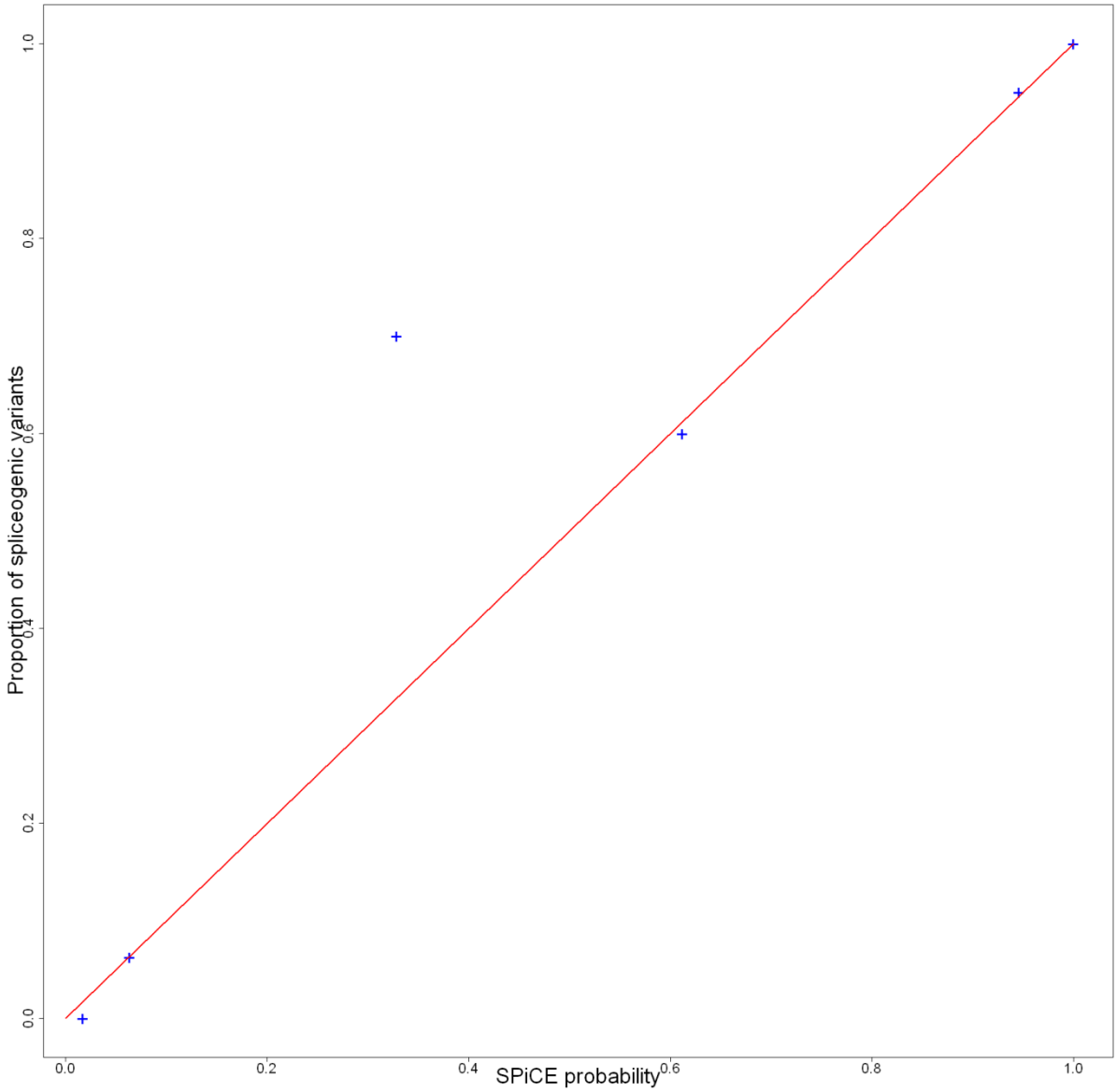


Figure S3: Proportion of spliceogenic variants according to probability to alter splicing on validation sets (n = 250). Blue points represent correlation between proportions of variants and probability of an effect on splicing, red line corresponds to y=x. Variants were subdividing in 6 groups according to their fitted probability: [0; 0.027], [0.027; 0.115], [0.115; 0.432], [0.432; 0.749], [0.749; 0.998] and [0.998; 1]. Effective for each group was 16, 16, 10, 15, 40 and 153 respectively. Each blue point represents a group of variants.

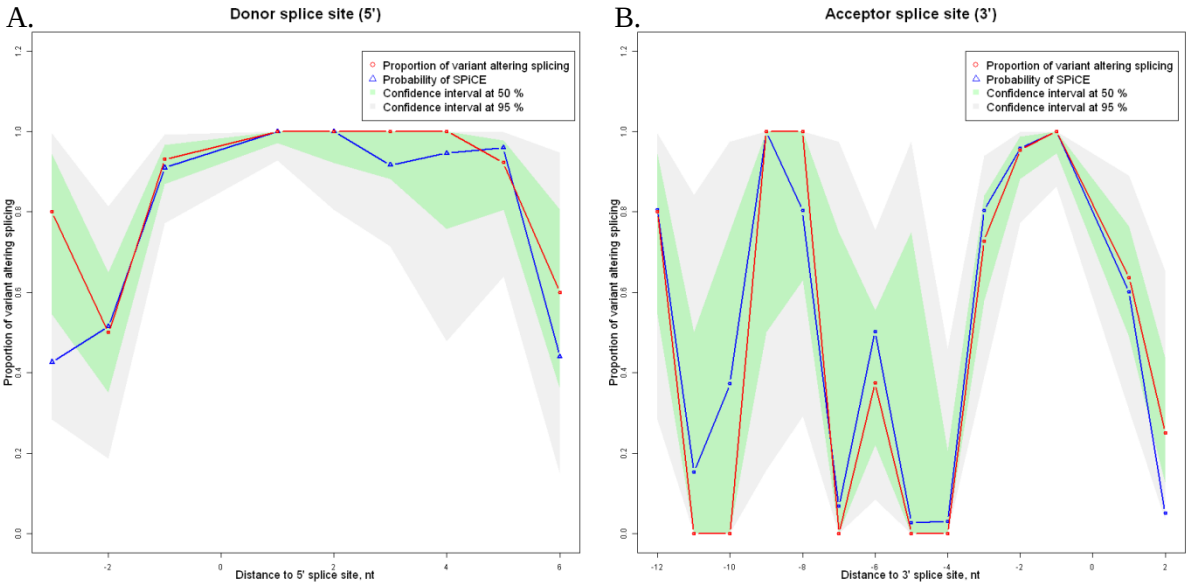


Figure S4: Comparison between the proportion of spliceogenic variants and the probability of SPiCE for each position on validation sets (n = 250). A) variants localized in donor splice site from -3 to 8. **B)** variants localized in acceptor splice site from -12 to 2. Red lines represent the observed proportion of spliceogenic variants for each position. Blue lines represent the average of SPiCE probability for each position. Green areas represents the confidence interval at 50% and grey areas the confidence interval at 95%.

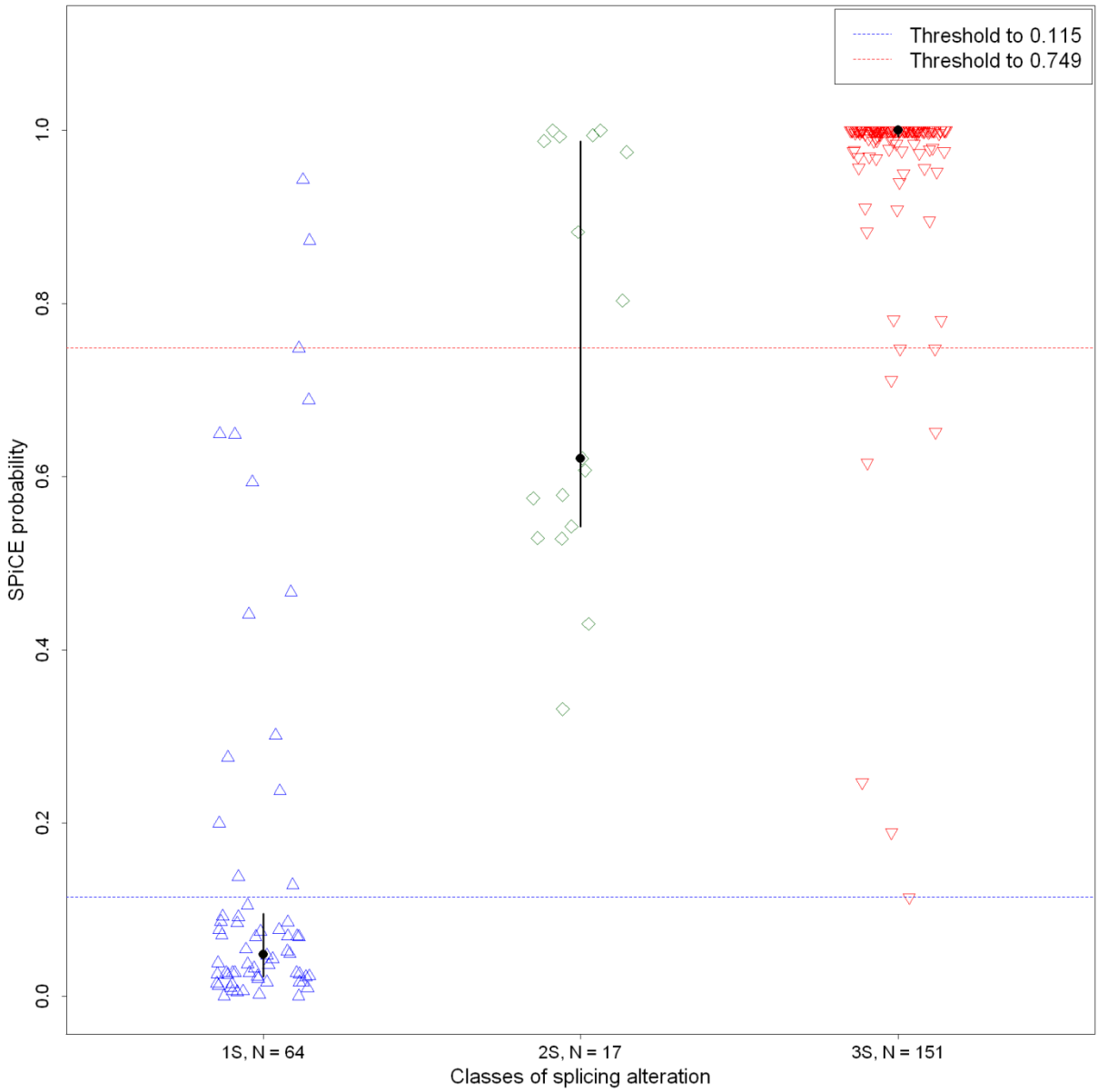
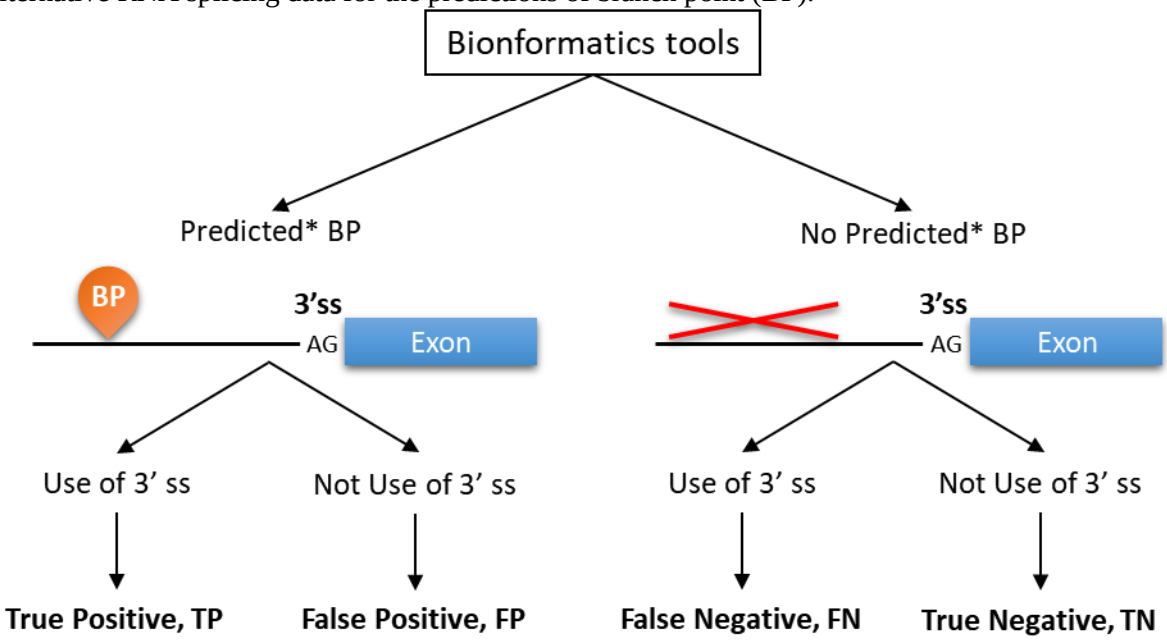


Figure S5: Predictive capacity of quantitative effects (n= 232). x-axis: the semi-quantitative nature denoted as 1S (no effect on splicing, n=64), 2S (partial effect, n=17) and 3S (complete effect, n=151) is plotted against SPiCE probabilities (y-axis). Optimal sensitivity threshold (ThSe, probability above 0.115) and optimal specificity threshold (ThSp, probability above 0.749) are indicated by dotted lines. Nine out of 17 partial effects are predicted within the medium range (0.115 to 0.749), see text for details.

II. ANNEXE B SUPPLEMENTARY INFORMATION: 'Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants'

The supplementary Table S1 and Table S2 are provided separately in Excel format, available at https://github.com/raphaelleman/BenchmarkBPprediction.

SUPPLEMENTARY Figure S1: Workflow to compare bioinformatics tools on physiological and alternative RNA splicing data for the predictions of branch point (BP).



	True 3' ss	False 3' ss
Predicted BP	TP	FP
No Predicted BP	FN	TN

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

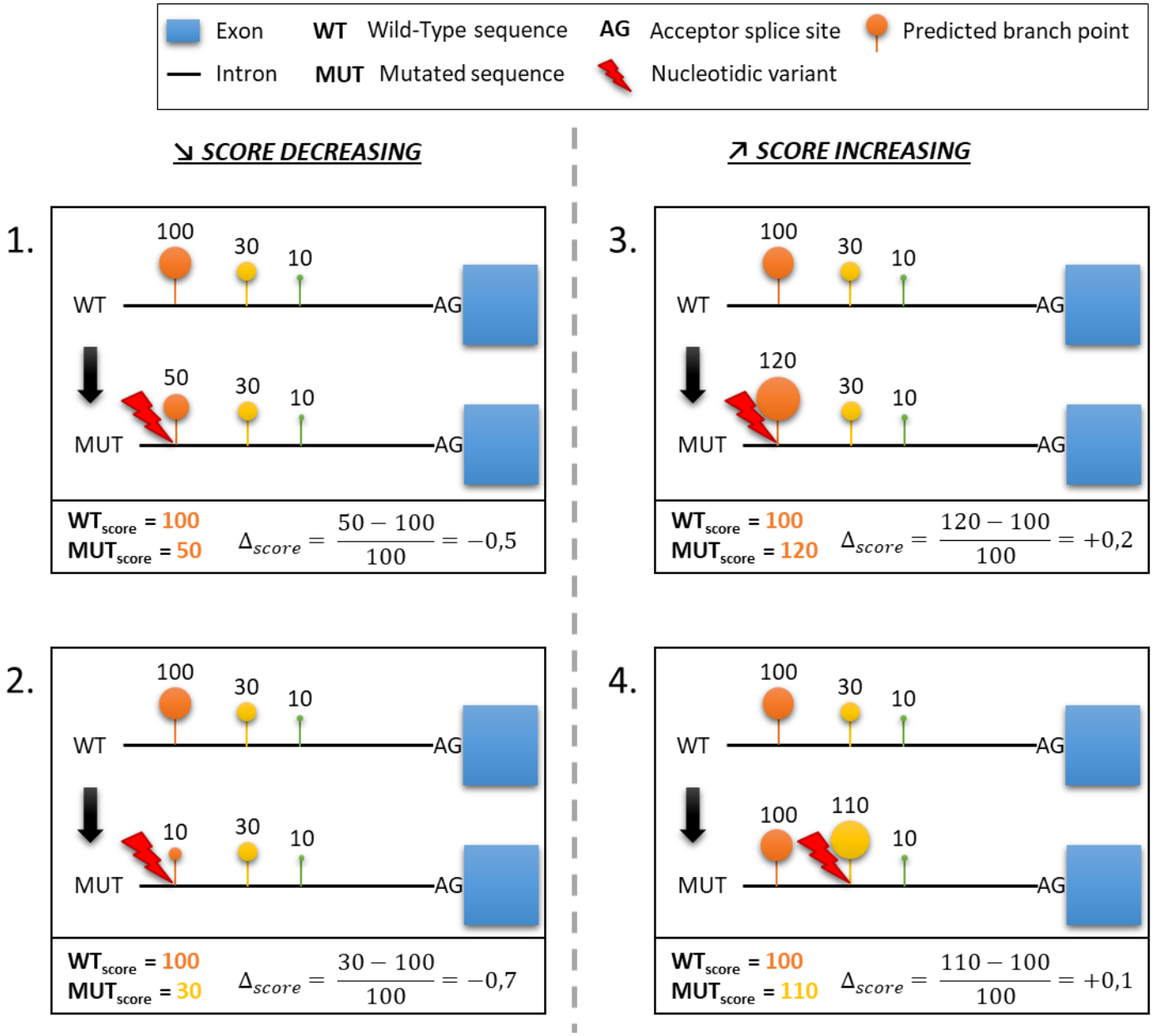
$$Specificity = \frac{TN}{TN + FP}$$

$$Positive\ Predictive\ Value, PPV = \frac{TP}{TP + FP}$$

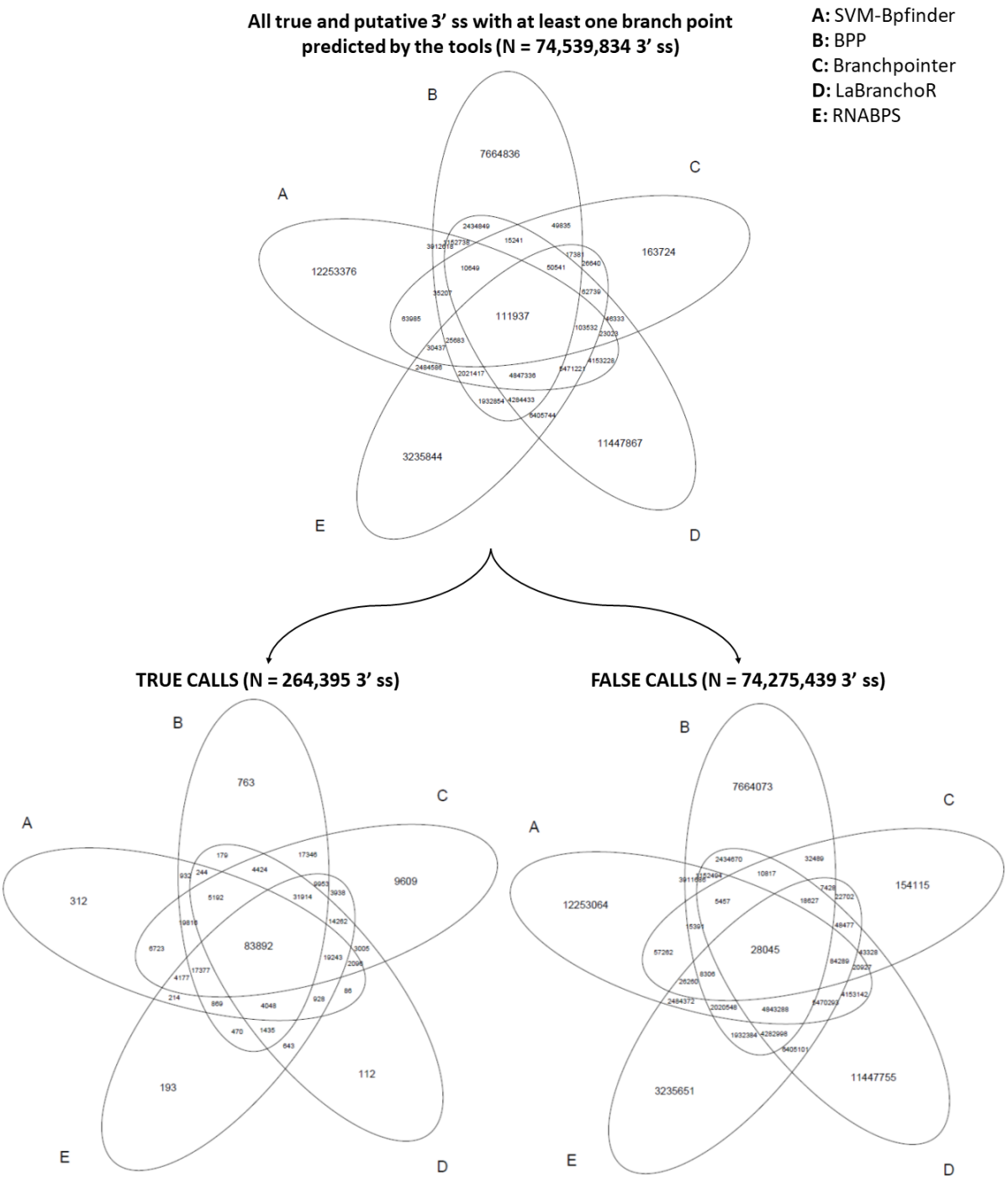
$$Negative\ Predictive\ Value, NPV = \frac{TN}{TN + FN}$$

*Prediction based on the optimal threshold obtained from the ROC analysis, excepted for Branchpointer that displays only BP with high confident level.

SUPPLEMENTARY Figure S2: The different ways that a variant may alter the branch point score.



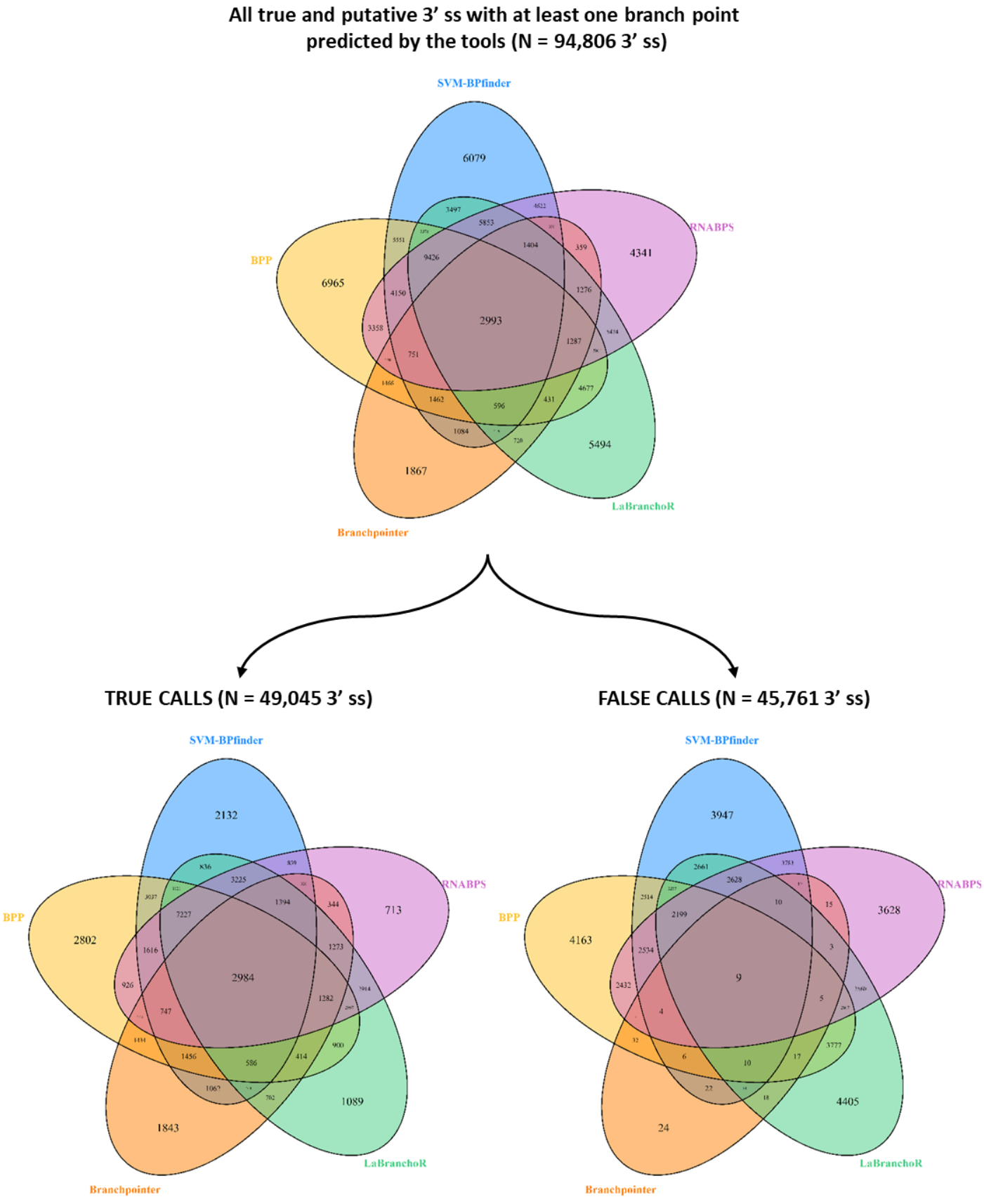
SUPPLEMENTARY Figure S3: The overlap of natural 3’ ss (True Calls) and controls AG (False Calls) from Ensembl data.



SUPPLEMENTAL Figure S3 (continuing):

Score combinaison	True calls	False calls	ratio True calls/all calls
SVM-BPfinder; BPP; Branchpointer; LaBranchoR; RNABPS	83892	28045	74.95%
SVM-BPfinder; BPP; Branchpointer; RNABPS	17377	8306	67.66%
BPP; Branchpointer; LaBranchoR; RNABPS	31914	18627	63.14%
BPP; Branchpointer; RNABPS	9953	7428	57.26%
SVM-BPfinder; BPP; Branchpointer	19816	15391	56.28%
SVM-BPfinder; BPP; Branchpointer; LaBranchoR	5192	5457	48.76%
BPP; Branchpointer	17346	32489	34.81%
BPP; Branchpointer; LaBranchoR	4424	10817	29.03%
Branchpointer; LaBranchoR; RNABPS	14262	48477	22.73%
SVM-BPfinder; Branchpointer; LaBranchoR; RNABPS	19243	84289	18.59%
Branchpointer; RNABPS	3938	22702	14.78%
SVM-BPfinder; Branchpointer; RNABPS	4177	26260	13.72%
SVM-BPfinder; Branchpointer	6723	57262	10.51%
SVM-BPfinder; Branchpointer; LaBranchoR	2096	20927	9.10%
Branchpointer; LaBranchoR	3005	43328	6.49%
Branchpointer	9609	154115	5.87%
SVM-BPfinder; BPP; LaBranchoR; RNABPS	4048	4843288	0.08%
SVM-BPfinder; BPP; RNABPS	869	2020548	0.04%
BPP; LaBranchoR; RNABPS	1435	4282998	0.03%
BPP; RNABPS	470	1932384	0.02%
SVM-BPfinder; BPP	932	3911686	0.02%
SVM-BPfinder; BPP; LaBranchoR	244	1152494	0.02%
SVM-BPfinder; LaBranchoR; RNABPS	928	5470293	0.02%
LaBranchoR; RNABPS	643	6405101	0.01%
BPP	763	7664073	0.01%
SVM-BPfinder; RNABPS	214	2484372	0.01%
BPP; LaBranchoR	179	2434670	0.01%
RNABPS	193	3235651	0.01%
SVM-BPfinder	312	12253064	0.00%
SVM-BPfinder; LaBranchoR	86	4153142	0.00%
LaBranchoR	112	11447755	0.00%

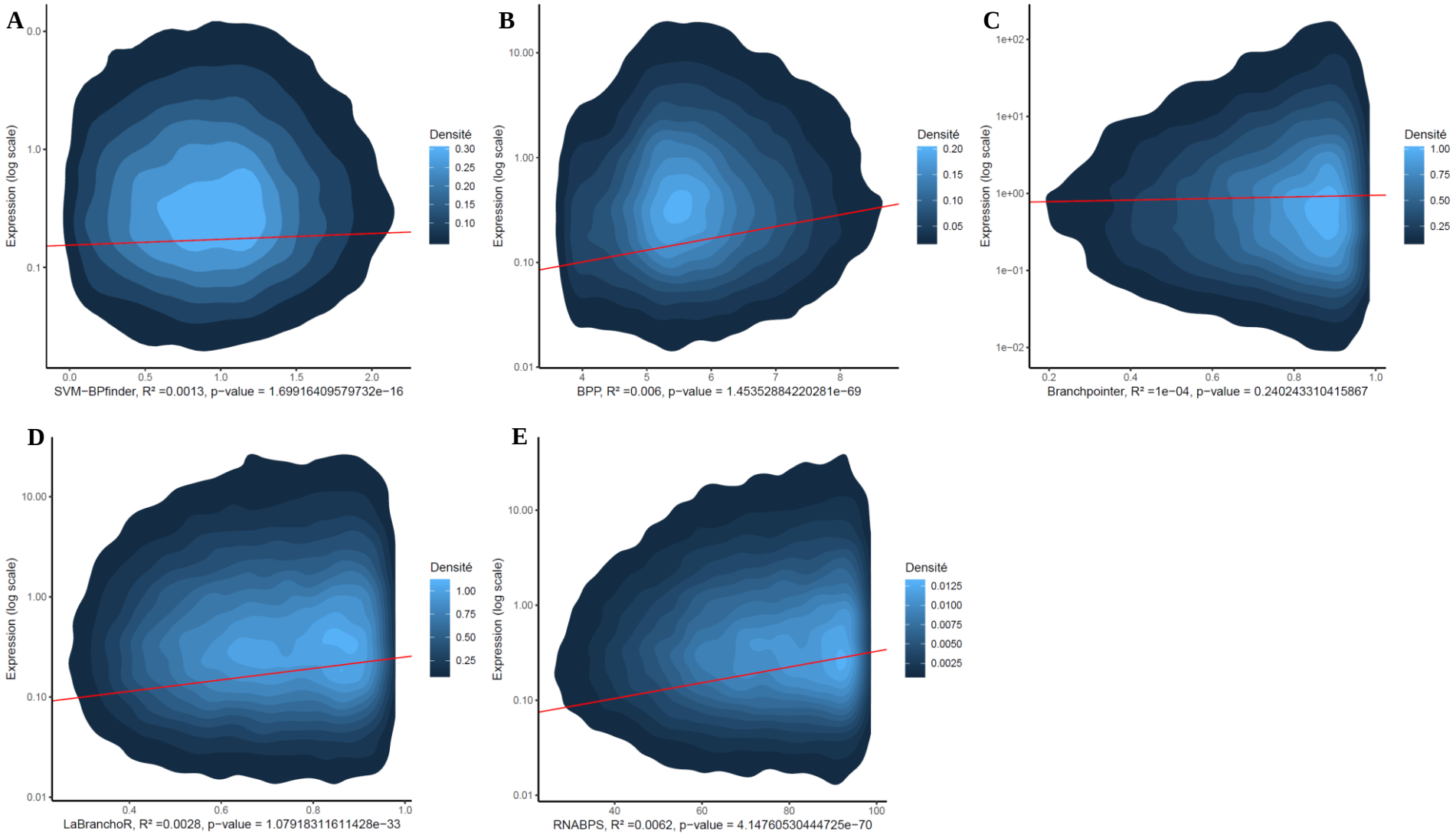
SUPPLEMENTARY Figure S4: The overlap of alternative 3’ ss (True Calls) and controls AG (False Calls) from our RNAseq data.



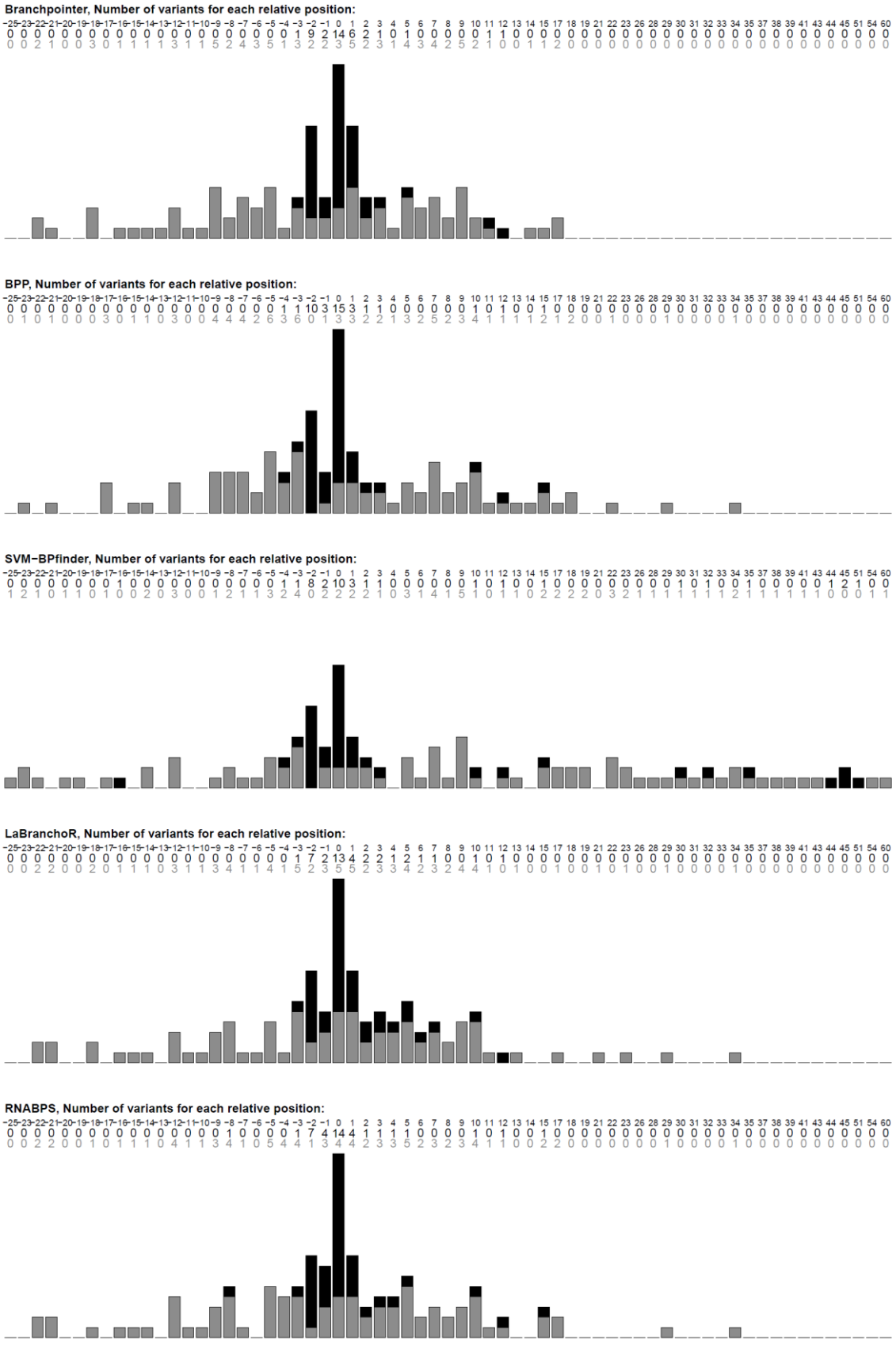
SUPPLEMENTAL Figure S4 (continuing):

Score combinaison	True calls	False calls	Ratio True calls/all calls
Branchpointer; LaBranchoR; RNABPS	1273	3	99,76%
SVM-BPfinder; BPP; Branchpointer; LaBranchoR; RNABPS	2984	9	99,70%
BPP; Branchpointer; LaBranchoR; RNABPS	1282	5	99,61%
SVM-BPfinder; BPP; Branchpointer	1456	6	99,59%
SVM-BPfinder; BPP; Branchpointer; RNABPS	747	4	99,47%
SVM-BPfinder; Branchpointer; LaBranchoR; RNABPS	1394	10	99,29%
Branchpointer	1843	24	98,71%
BPP; Branchpointer; RNABPS	423	7	98,37%
SVM-BPfinder; BPP; Branchpointer; LaBranchoR	586	10	98,32%
SVM-BPfinder; Branchpointer	1062	22	97,97%
BPP; Branchpointer	1434	32	97,82%
Branchpointer; LaBranchoR	702	18	97,50%
SVM-BPfinder; Branchpointer; RNABPS	326	10	97,02%
SVM-BPfinder; Branchpointer; LaBranchoR	401	14	96,63%
BPP; Branchpointer; LaBranchoR	414	17	96,06%
Branchpointer; RNABPS	344	15	95,82%
SVM-BPfinder; BPP; LaBranchoR; RNABPS	7227	2199	76,67%
BPP; LaBranchoR; RNABPS	2997	2067	59,18%
SVM-BPfinder; LaBranchoR; RNABPS	3225	2628	55,10%
SVM-BPfinder; BPP	3037	2514	54,71%
LaBranchoR; RNABPS	2914	2560	53,23%
BPP	2802	4163	40,23%
SVM-BPfinder; BPP; RNABPS	1616	2534	38,94%
SVM-BPfinder	2132	3947	35,07%
SVM-BPfinder; BPP; LaBranchoR	1121	2257	33,19%
BPP; RNABPS	926	2432	27,58%
SVM-BPfinder; LaBranchoR	836	2661	23,91%
LaBranchoR	1089	4405	19,82%
BPP; LaBranchoR	900	3777	19,24%
SVM-BPfinder; RNABPS	839	3783	18,15%
RNABPS	713	3628	16,42%

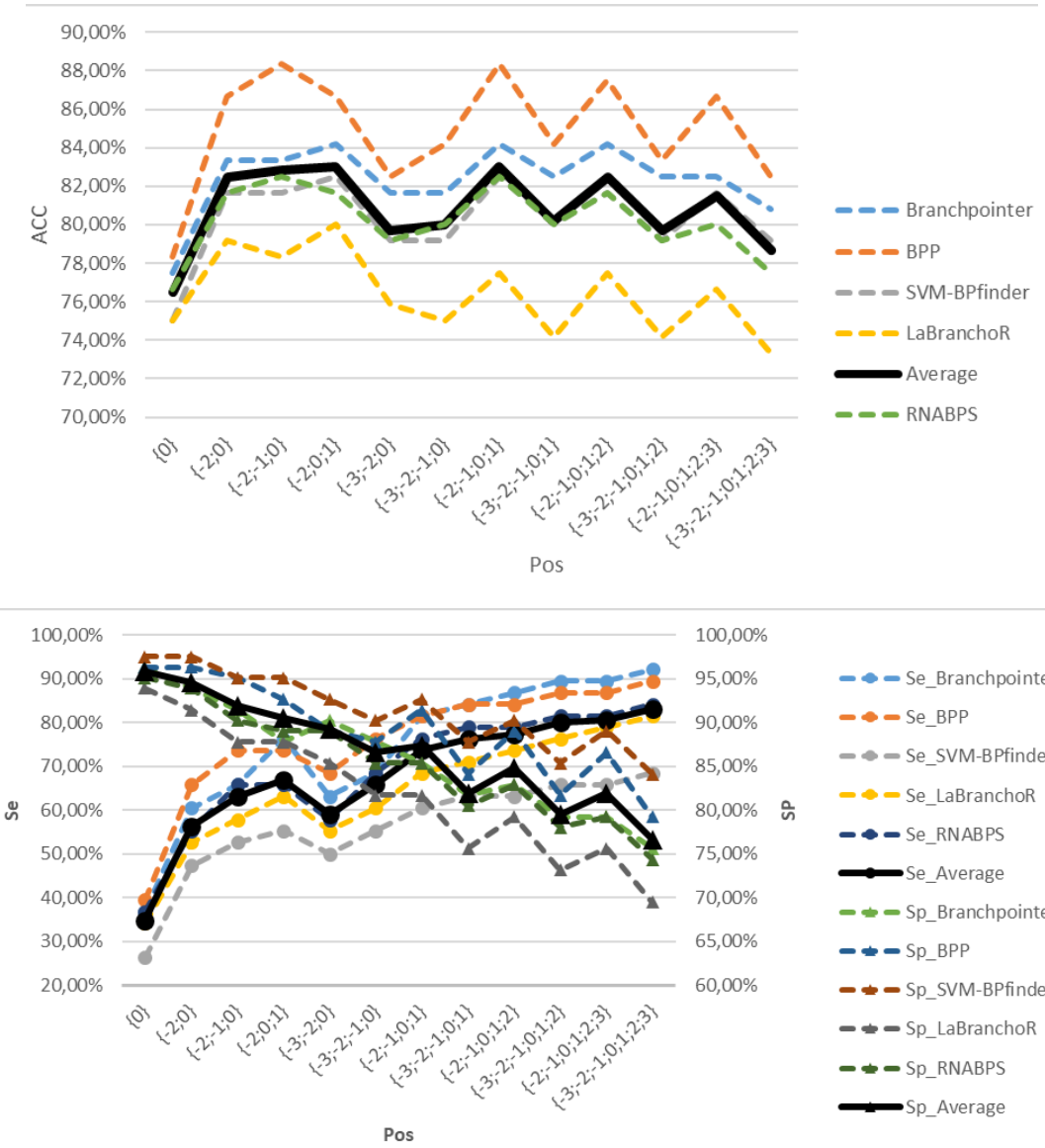
SUPPLEMENTARY Figure S5: Correlation between the scores (SVM-BPfinder, BPP, Branchpointer, LaBranchoR, RNABPS) and the expression of alternative 3'ss. **A**: SVM-BPfinder (R^2 = 0.0013, p-value = 1.70x10-16), **B**: BPP (R^2 = 0.006, p-value = 1.45x10-69), **C**: Branchpointer (R^2 = 0.0001, p-value = 0.24), **D**: LaBranchoR (R^2 = 0.0028, p-value = 1.08x10-33), **E**: RNABPS (R^2 = 0.0062, p-value = 4.14x10-70).



SUPPLEMENTARY Figure S6: Repartition of variants (n = 120) according their position relative to the predicted branch point. Position 0 corresponds to the branch point adenosine. Shown in black are variants affecting splicing, and in grey the variants without effect on mRNA splicing.

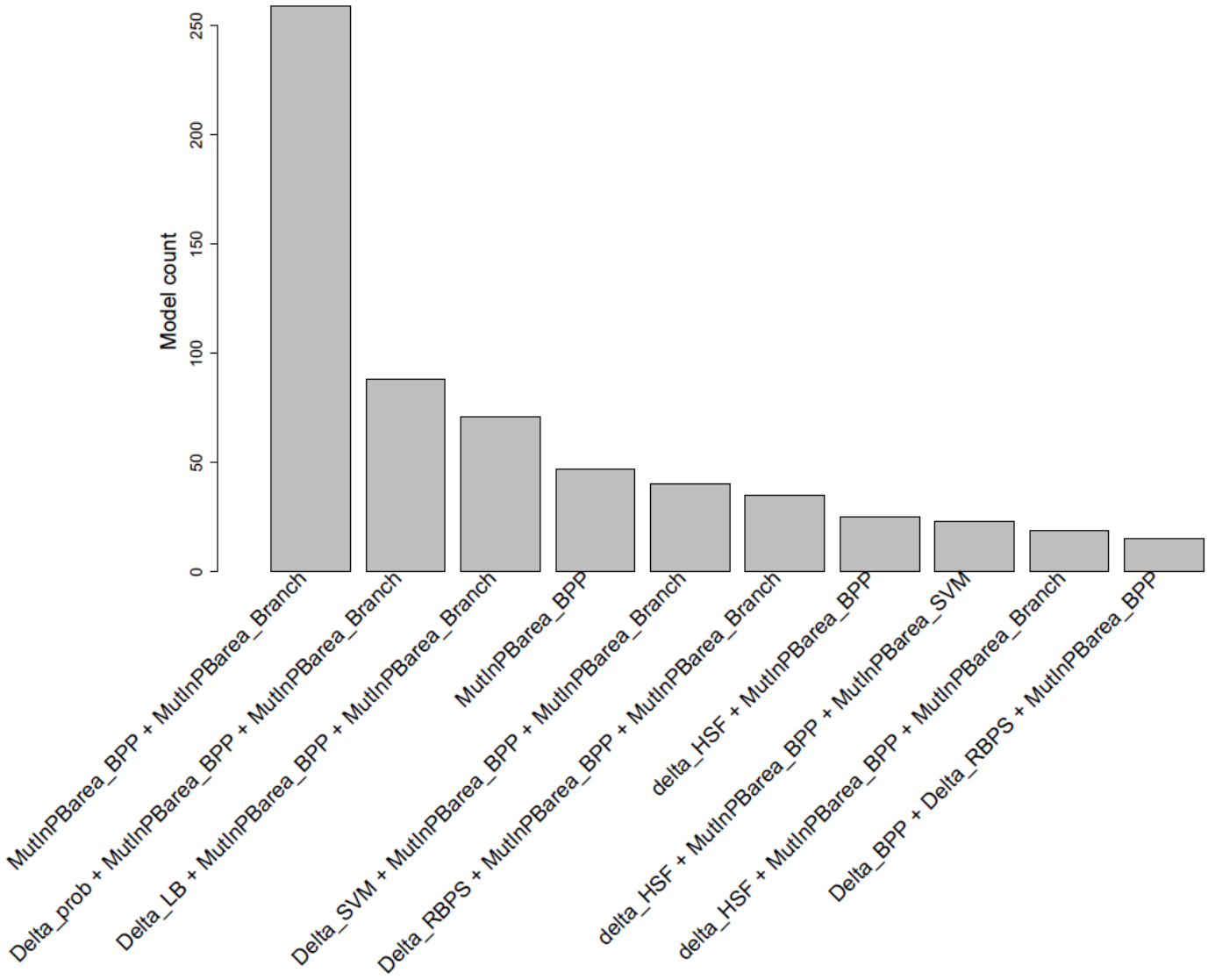


SUPPLEMENTARY Figure S7: Determination of optimal motif (YTRAYNN) length to predict splicing alteration, n = 120 variants. ACC: Accuracy, Pos: relative position in branch point motif, Se: Sensitivity, Sp: Specificity.



Relative position in the motif	Sequence motif (7-mer: YTRAYNN)
{0}	A
{-2;0}	T-A
{-2;-1;0}	TRA
{-2;0;1}	T-AY
{-3;-2;0}	YT-A
{-3;-2;-1;0}	YTRA
{-2;-1;0;1}	TRAY
{-3;-2;-1;0;1}	YTRAY
{-2;-1;0;1;2}	TRAYN
{-3;-2;-1;0;1;2}	YTRAYN
{-2;-1;0;1;2;3}	TRAYNN
{-3;-2;-1;0;1;2;3}	YTRAYNN

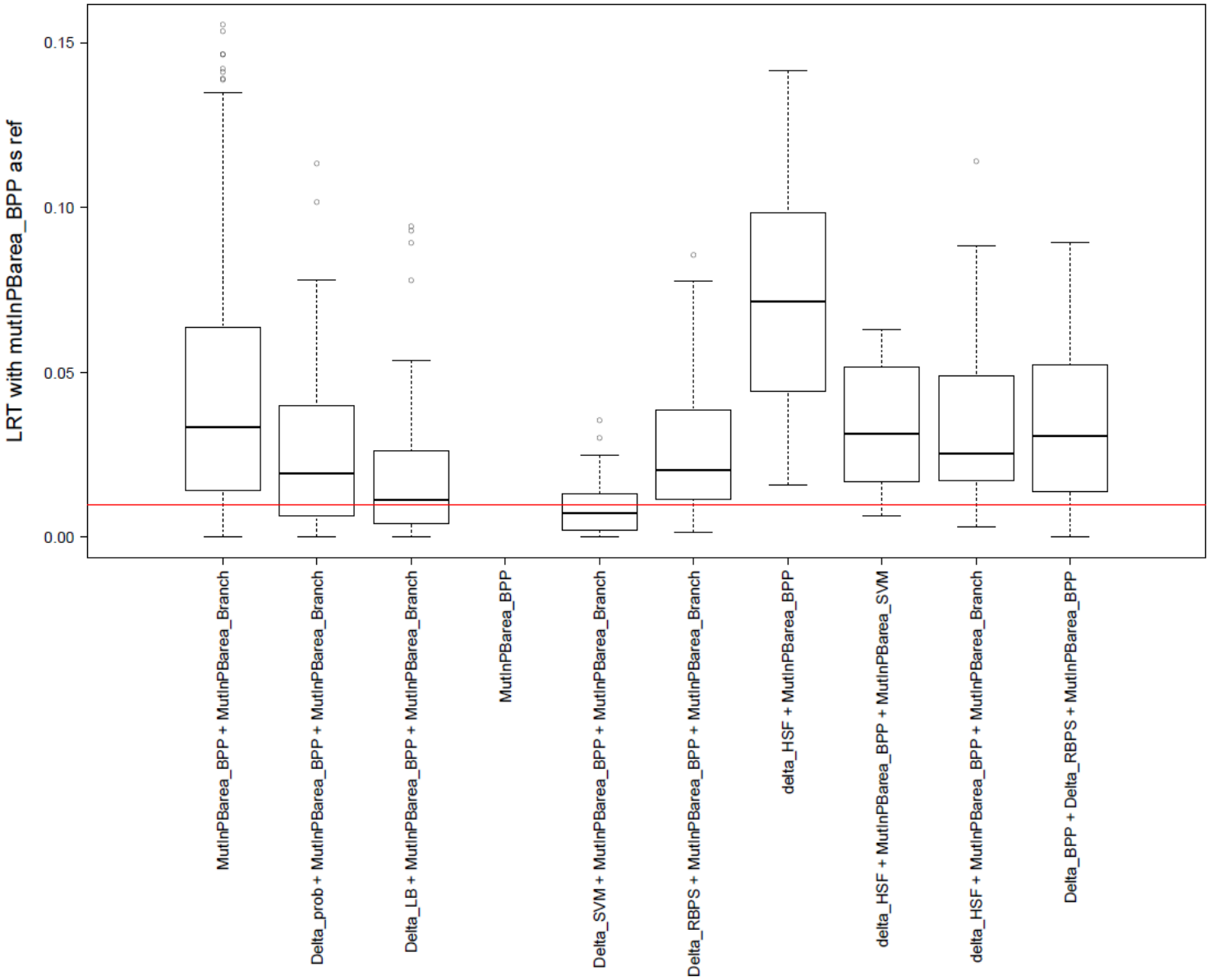
SUPPLEMENTARY Figure S8: Cross-validation (1,000 times) to select the optimal model to predict branch point alteration.



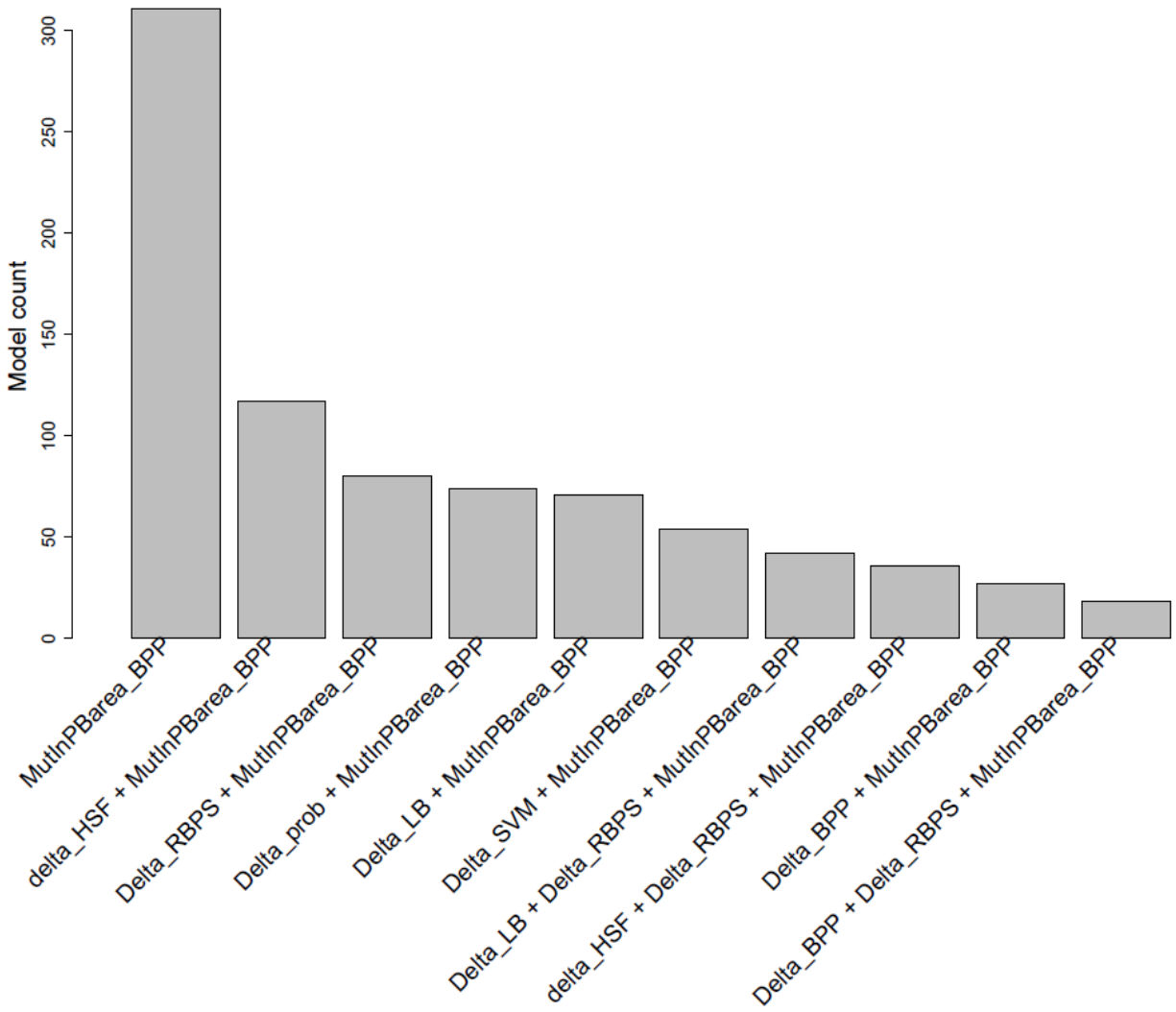
Acronym of variables	Explanation
Delta	Variation score between wild Type-mutated
MutInPBarea	Is variant located in the 4-mer of BP-predicted

Acronym of scores	Complete score names
HSF	Human Splicing Finder
SVM	SVM-BPfinder
BPP	Branch Point Predictor
Branch	Branchpointer
LB	LaBranchoR
RNABPS	RNA Branch Point Selection

SUPPLEMENTARY Figure S8 (continuation): Likelihood ratio test (LRT) between univariate model and tested model. The univariate model was position of predicted-BP by BPP alone. The red line represents the p-value of 1 %.



SUPPLEMENTARY Figure S9: Cross-validation (1,000 times) to select the optimal model to predict branch point alteration without the positions of predicted BP for all tools except BPP.



Acronym of variables	Explanation
Delta	Variation score between wild Type-mutated
MutInPBarea	Is variant located in the 4-mer of BP-predicted

Acronym of scores	Complete score names
HSF	Human Splicing Finder
SVM	SVM-BPfinder
BPP	Branch Point Predictor
Branch	Branchpointer
LB	LaBranchoR
RNABPS	RNA Branch Point Selection

III. ANNEXE C: SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations, validated on a curated diagnostic set of 2,784 exonic and intronic variants.

1. Main text

INTRODUCTION

Splicing alterations are implicated in a large variety of disease phenotypes and are presumably the most frequent alterations involved in hereditary disease [1], [2]. The reason is that each nucleotide variation, regardless of its location, can potentially impact on splicing. Recent data showed that close to 4% of Exac variants lead to a splice alteration[3], making their detection mandatory for the genetic diagnosis of hereditary and somatic diseases such as cancer, and more broadly for optimized genomic medicine. Splice alterations are highly diverse in nature e.g. single or multi (cassette) exon skipping, use of cryptic splice site and splice site shifting with following exonic deletion or intronic retention/pseudo exon [4]. Pre-mRNA splicing requires a number of dedicated sequences, the splice donor site (5'ss), the splice acceptor site (3'ss), the branch point (BP), the polypyrimidine tract (PPT) located between the BP and the 3'ss, and complementary motifs called splicing regulatory elements (SREs) located close to the 5'/3' splice sites (Supplementary Figure S1). The diversity of splicing alterations results from the disruption and/or creation of one or more of these splicing consensus elements. Consequently, most *in silico* prediction tools are dedicated to these specific splicing motifs and don't provide a comprehensive assessment of the whole gene sequence.

To provide the community with a comprehensive prediction tool dealing with the diversity of splice alterations, our international consortium gathered a curated set of 2,784 variants in 213 genes with their corresponding splicing studies.

This dataset was used to assay the Splicing Prediction Pipeline (SPiP). SPiP is an application suite, running a cascade of optimal and complementary bioinformatics tools. The definition of optimal tools was made by a review of literature. If for particular motif the optimal was not defined, then we assayed a set of tools and their combination. The evaluation of overall SPiP performances was performed on this evaluation set of 2,784 variants.

SPiP detected all classes of splice alterations (exon skipping, use of new splice site, intronic exonisation and intronic retention) with high sensitivities ranging from 81.25 % to 96.03%.

As this curated data set doesn't represent the diagnostic situations geneticists have to face, and in order to mimic routine diagnostic use, we added 45 000 variants with MAF>5% and run SPiP. SPiP reached an accuracy of 80.76 %, a sensitivity of 97.81 % and a specificity of 80.46%.

RESULTS

1. Collection of variants with their *in vitro* RNA studies

The evaluation set of 2,784 distinct variants from 213 genes comprised 47 % (N = 1,294) of spliceogenic variants dysplaying all kind of alterations: exon skipping (N = 831, 64.2 %), splice site shifting (N = 359, 27.7 %), intronic exonisation (N = 88, 6.8 %) and intronic retention (N = 16, 1.2 %) (Table 1 and Figure 1**Figure 1:** Repartition of variants in the different splicing motif, N = 2,784 variants. BP: Branch point area, PolyTC: polypyrimidine tract, Cons: Consensus splice site.).

Table 1: Splicing alteration observed in the 2,784 variants

Impact on splicing	N (%)
Without impact on splicing	1490 (53 %)
Splicing alteration	1294 (47 %)
Exon skipping	831 (64 %)
Splice site shift	359 (28 %)
Pseudo-exon	88 (7 %)
Intron retention	16 (1 %)

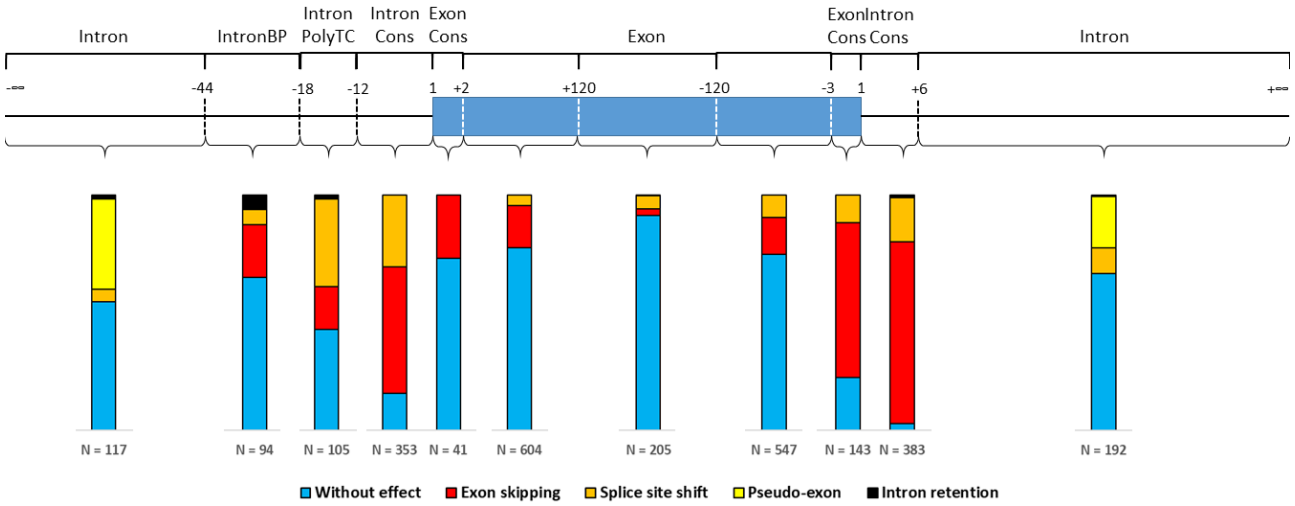


Figure 1: Repartition of variants in the different splicing motif, N = 2,784 variants. BP: Branch point area, PolyTC: polypyrimidine tract, Cons: Consensus splice site.

Among the 359 variants causing splice site shift, 195 variants created a new splice site and 164 variants altered the 5' or 3' wild type natural splice site. We observed that the splicing impact depended on the distance of the spliceogenic variant to the WT splice site. A *de novo* splice site can emerged if the variant is located within a 653 bp-window from the wild type splice site, with a 55 bp median distance. On the other hand, cryptic splice sites could be used regardless of the distance to the wild type splice site, with a median distance of 49 bp. However, variants creating new splice site were near to de novo/cryptic splice sites, 99 % of variants were at least 35 nt of new splice site. The median distance was one nucleotide, i.e the variants affected the canonical splice site.

Intronic variants can create de novo splice sites (N = 62 variants) or pseudo-exons (N= 87 variants). Similarly, we found that the distance between the intronic variant and the wild type splice site drives the impact as 95.2 % (59/62) of variants before 150 bp led to a splice shift whereas 94.2 % (82/87) of variants occurring beyond 150 bp from the natural splice site led to pseudo-exonisation.

2. Optimal bioinformatic tools

To build spip, the best tools were selected across the different splice motifs and using subsets of the data collection.

a. Consensus splice sites

For consensus 5' and 3' splice sites, the combination of MES and SSF tools was defined to be optimal [5], [6]. And recently, the tool SPiCE was published and gathers the score of MES and SSF [7]. Indeed, SPiCE had shown best performance on the validation set than MES and SSF alone. Adding to this, we compared SPiCE with ADAboost, MES, SPANR on a subset of 253 variants used as SPiCE validation set (Supplementary Figure S5). As demonstrated by ROC curve analyses, SPiCE showed the best performance (AUC of 0.978 for SPiCE, 0.971 for ADAboost and MES, 0.891 for SPANR). SPiCE provide also the decision making thresholds to predict or not splicing alteration.

b. Polypyrimidine tract (PPT)

For the PPT, we used the consensus prediction tools extending their prediction up to the branch point area. This one starts at the 18th nt in intron. Among optimal consensus tools MES and SSF, only MES reaches this border. To confirm the efficiency of MES to predict splicing alteration in this motif, we had to use the 63 variants occurred between -13 and -17 in the set of 2,784 variants. We compared the performances of MES on these 63 variants with the deep learning tool, SPANR. MES outperformed SPANR as shown by ROC curve analyses (AUC 0.909 vs 0.639, see Supplementary Figure S6). The decision making threshold of MES was set to 15 % [5].

c. Branch point (BP)

For BP alterations, we used our previous work to compare the tool dedicated to BPs [8]. Indeed, we have compared a set of recently published tools. Thus, we found that the optimal prediction strategy should not rely on score calculations, but rather on the presence/absence of the variant in the 4-mers motif (TRAY) of the predicted branchpoint. BPP showed the best performance with an accuracy of 89.17 % vs 81.67 for BP finder (Supplementary Table S3).

d. Exonic Splicing Regulators (ESR)

The tool ΔtESRseq was defined as one of the optimal tools for ESR prediction but previously lacked a decision threshold [9]–[11]. Therefore, we had to set this threshold and to reduce the risk or overfitting

we processed to a cross-validation. This cross-validation was 1,000 iters on 1,068 variants from the set of 2,784 variants. The median threshold of -1.10 (sd: 0.356) was determined, providing a 70.21 % accuracy, 67.65 % sensitivity and 74.65 specificity.

e. Splice site creation or reinforcement (5'ss/3'ss de novo/cryptic)

Due to the lack of benchmarks and guidelines to predict the creation (*i.e. de novo*) or reinforcement (*i.e.* cryptic) of splice site by a variant, we developed a metascore model using 202,989,656 splice sites, 530,931 effective splice sites and 202,458,725 AG/GT controls from Ensembl [12]. Training phase on 135,326,351 splice sites showed that the combination of three scores (MES, PWM and ESR scores) significantly improved the model (p-value of Wald test < 10^{-7}). The validation phase on 67,663,176 splice sites reached an area under the ROC curve of 0.974 and an overall accuracy of 92.7 % using an optimal threshold of 0.039 (Figure 2). The correspondence between the proportion of splice used and the model probability was also confirmed until to probability of 0.7 (Supplementary Figure S7). Indeed, beyond 0.7, the weak number of splice sites did not permit to conclude. Then we proposed the strategy described in Figure 3 to detect *de novo*/cryptic splice sites. During our study, SpliceRover was published as a new tool to detect cryptic splice site with high performance [13]. On our data set, we can observe a better accuracy of SPiP (84.6 %) compare to Spice Rover (75.2 %).

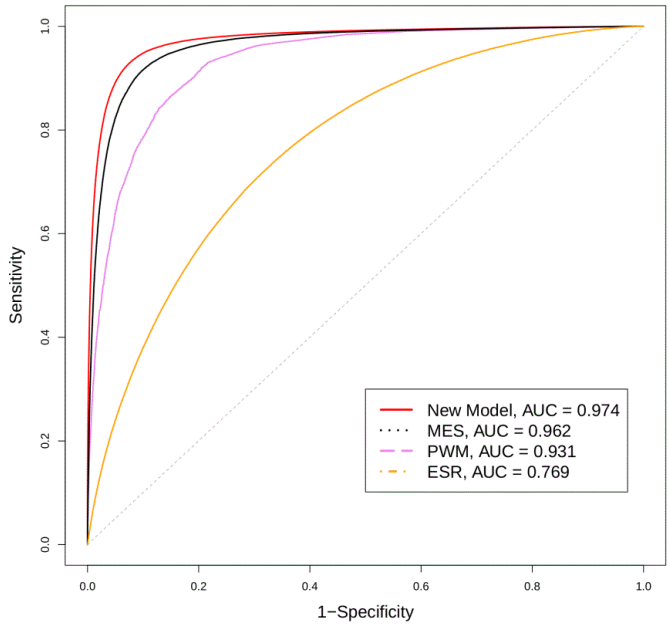


Figure 2: Performance of the new model to predict splice site used versus the each individual scores: MaxEntScan (MES), Position Weight Matrix (PWM), Exonic Splicing Regulator scores (ESR). ROC curves performed on validation set N = 67,663,176 splice sites.

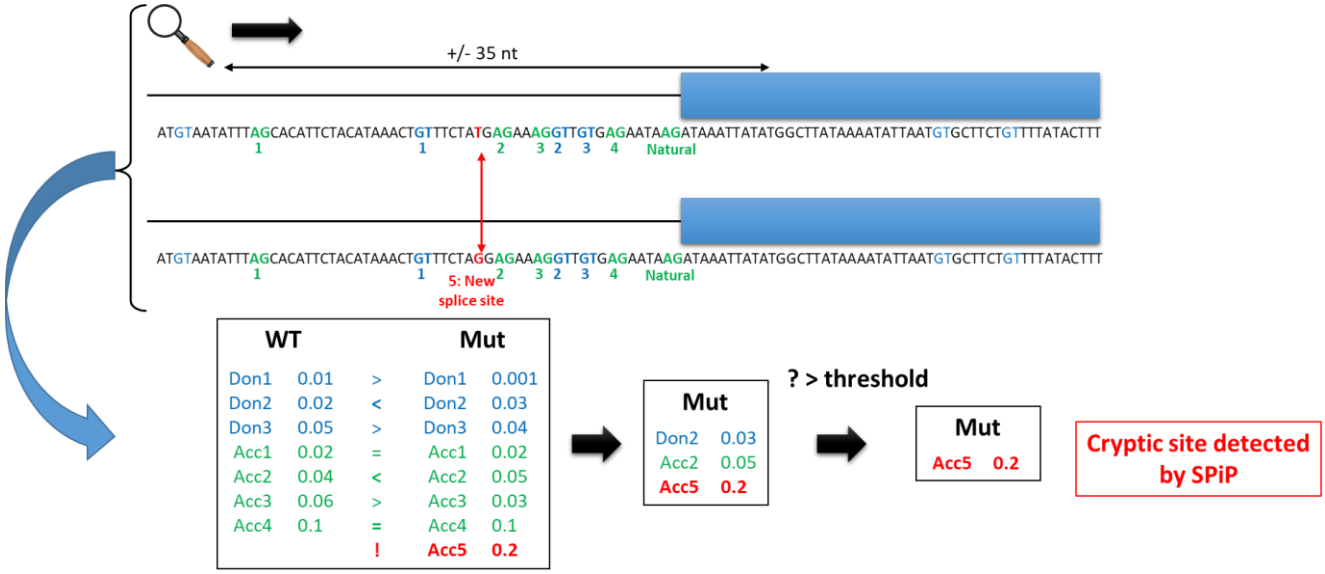


Figure 3: The strategy used by our tool to detect an activation of cryptic splice site by a mutation. In the scan area the tool detects 4 AG signals and 3 GT signals in wild-type and mutated sequences, plus a fifth *de novo* AG signal in mutated sequence, (*i.e.* potential splice sites). The tool compares the score of each potential splice sites. Only the second donor site and the second acceptor sites have a reinforcement of score plus the *de novo* acceptor site (Acc5). On these 3 splice sites, only the *de novo* splice site has a score above the decisional threshold (see text) and so is predicted as *de novo*/cryptic splice site.

3. Overall performance of SPiP on 2,784 variants

Following this evaluation phase, SPiCE, MES, BPP, ΔtESRseq and the new metascore were embedded into SPiP to scan the entire nucleotidic/genomic context of a gene (Figure 4).

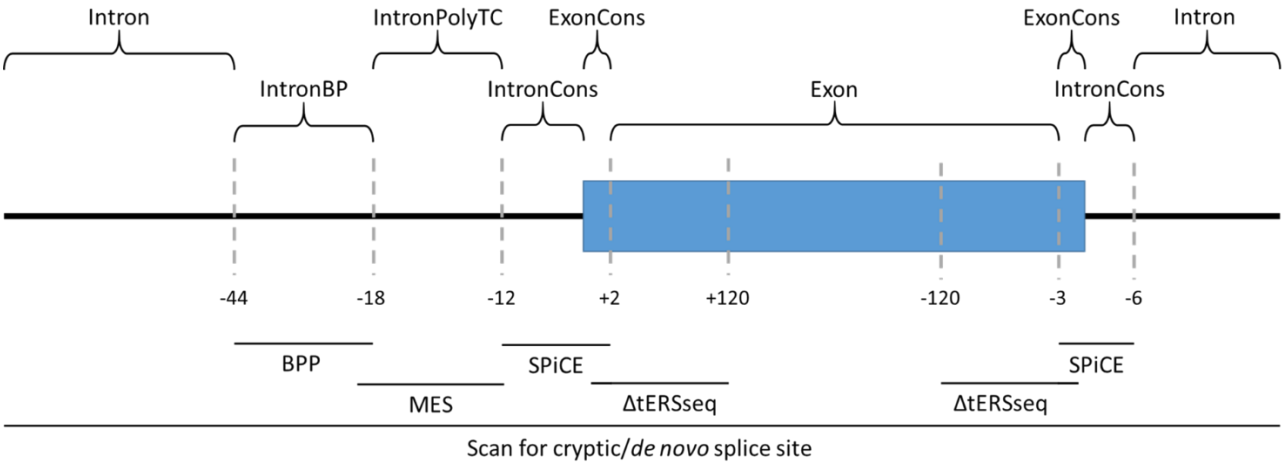


Figure 4: The bioinformatics tools used for each splicing motifs. IntronBP: Branch point area, IntronPolyTC: polypyrimidine tract, IntronCons: Intronic consensus splice site, ExonCons: Exonic consensus splice site.

On the overall set of 2784 variants, SPiP reached an accuracy of 80.21%, a sensitivity of 90.96 % and a specificity of 70.87 % (Table 2). With regards to the splicing defects, new splice site, exon skipping,

pseudo exon and intronic retention were detected with a sensitivity of 96.03 %; 89.77 %; 82.96 % and 81.25 %, respectively. The lower values obtained for pseudo exon and retention are probably due to their low representation among the 2784 variants (88 and 16 events on 1,294). We compared the SPiP performance with two deep learning algorithm SPIDEX [14] and SpliceAI [15] (Table 3). SPiP outperforms SPIDEX (accuracy of 80.21 % vs 75.45 %). SpliceAI shown high specificity (98.26 % vs 70.87 %). However, SpliceAI missed several spliceogenic variant with sensitivity of 70.71 % vs 90.96 %. Therefore, SPiP confirmed its interest to detect the presence of spliceogenic variants.

Table 2: Overall performances of SPiP on 2,784 variants. TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

	SPiP prediction
TP	1177
FP	434
TN	1056
FN	117
Accuracy	80.21%
Sensitivity	90.96 %
Specificity	70.87 %

Table 3: Performances of SPiP versus SPIDEX and SpliceAI on the 2,784 variants reported in our collection. TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative). [a]751 SPIDEX missing data due to delins variants and variants outside 300 nt in intron, 240 of them alter splicing

	SPiP prediction	SPIDEX[a]	SpliceAI
Accuracy	80.21%	75.45 %	85.45 %
Sensitivity	90.96 %	78.37 %	70.71 %
Specificity	70.87 %	72.32 %	98.26 %

4. SPiP and the probability of splicing alteration

SPiP shows remarkably high sensitivity values. However, these values were obtained on a selected set of data (1,294 spliceogenic variants among a total of 2,784 variants) and should be corrected to mimic real life genomic experiments. To this aim, we reasoned that the proportion of spliceogenic variants is similar to this real life: 414 among 1540 exonic variants impacted splicing *i.e.* 26.88 % ($CI_{95\%}$ [24.67 % – 29.09 %]), in accordance with the literature, 23.5 % ($CI_{95\%}$ [10.3% – 48.0 %]. We downloaded from the 1000 Genome project (download April 4, 2019), 37,939,863 intragenic variants, of whom 1,899,246 (5.01%) were intragenic [16]. As a result, the Bayesian probability that a variant alters splicing whatever its position in a gene was 2.69 % ($CI_{95\%}$ [2.57 % – 2.82 %]). This value seems to be in agreement with the latest estimate of 3.8% from the last high throughput splice assay on ExAC data [3]. As we had 1294 spliceogenic variants among our 2784 collection, we need to increase the dataset to 45,000 variants to be in line with this 2.69% value. Consequently, 55 487 variants (MAF>5%) from our 213 genes were extracted from UCSC (download March 11, 2019) [17], of whom 45 000 were randomly picked up and

added to the evaluation set, thus enabling a 2.71% of spliceogenic variants in this “real life” evaluation set.

On this new set of data (n = 47,784 variants), SPiP reached an accuracy of 80.76 %, sensitivity of 97.81 % and specificity of 80.46%. The highest estimated probability of splicing alteration, among SPiP alteration (positive predictive value, PPV), was for the motives: 5/3’ss, polypyrimidine tract and BPs. The lowest estimated probability was encounter for deep intronic variants. Importantly, the negative predictive value (NPV) was above 90 % regardless of variant location (Figure 5). The detailed probabilities of splicing alteration according to SPiP predictions and variant localization were illustrated in supplemental file (Figure S8).

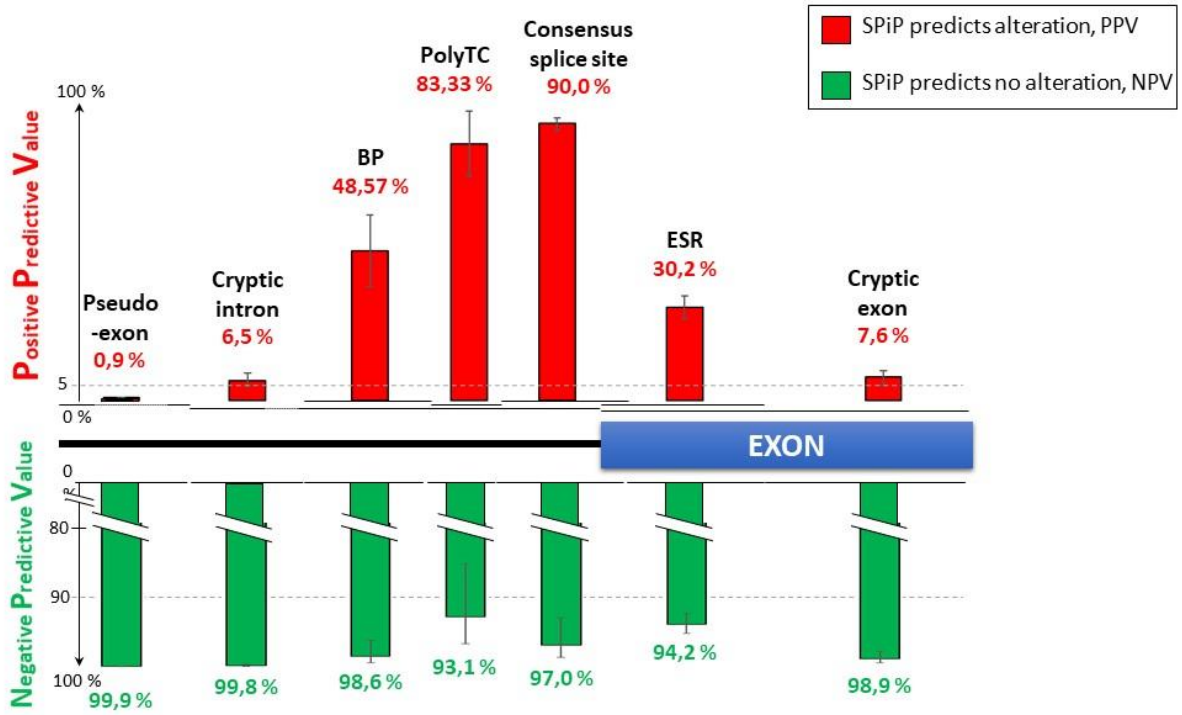


Figure 5: probability of splicing alteration according to SPiP prediction and variant localization in transcript. The detailed probabilities are shown in supplemental data.

5. SPiP performance on pathogenic variants

To estimate the capability of SPiP to detect pathogenic variants on another independent set of data, we dowloaded 2,936 variants from the ClinVar dataset (download May 24, 2019), and selected 317 (10.8 %) pathogenic or likely pathogenic spliceogenic variants [18]. SPiP reached an accuracy of 79.56%, a specificity of 77.24 % and a sensitivity of 98.74% (Supplementary Table S4).

DISCUSSION

SPiP is a splicing prediction suite addressing the different splicing motifs hence dealing with any variant whatever its location on the gene. It runs a combination of tools selected for their high level of

performance and consequently provides an unprecedented sensitivity of 90.96 % compared to recent deep learning prediction tools e.g. the popular SPIDEX (cited more than 600 times) and SpliceAI (Table 3). In order to facilitate interpretation, SPiP reports i) the altered consensus motif(s), unlike VEP from Ensembl [19], Spliceogen [20] or SpliceAI ii) the probability of splice alteration with the relevant negative and positive predictive values. Combination of both allows an easy and comprehensive interpretation of the results by the users.

To ensure an access of SPiP we proposed two versions of this tool for Windows and Linux OS. The Windows version offers a user-friendly interface (Figure 6) and the Linux version permits the parallel processing of a great number of variants. At last, parallel processing gave a ~0.5 sec/variant/CPU runtime under Linux CentOS (Supplementary Figure S9).

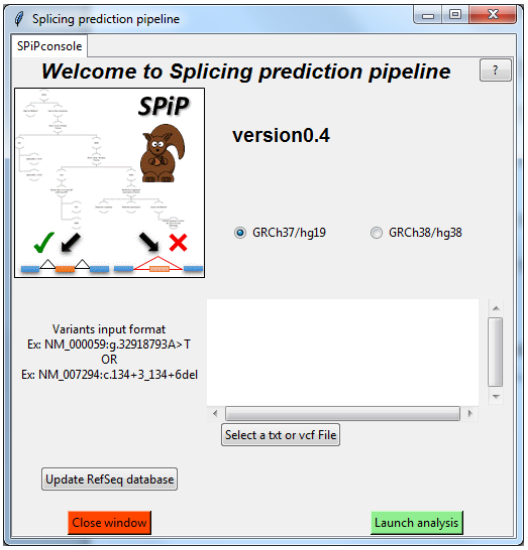


Figure 6: Console of SPiP under Windows version.

In this era of big data for genomic medicine, both PPV and NPV are important. In other words, detecting true spliceogenic variants is needed but discarding the bunch of true negatives is almost also important. PPV and NPV vary according to the consensus motifs (Figure 5) hence the location on the genomic sequence which is why the altered motif is indicated in SPiP output. PPV varies from 0.9% to 90% (deep intronic and consensus 5'- 3', respectively), meaning predictions of true spliceogenic variants are poor deep in introns and excellent within consensus sites. Positive predictions are also reliable within the polypyrimidine tract (83.33%) and mitigated for variants occurring in branch points (48.57%). Caution is needed for deep intronic and exonic variants as a maximum PPV of 30.2% is obtained for ESR motifs and pseudo exon and de novo intronic are detected with a PPV of 0,9% and 6.5%, respectively. The reason is that for ESR and cryptic motives, predictions are contaminated by a high false-positive rate. This can be explained by the low conservation of ESR motives and by the huge number of potential cryptic sites

On the other hand, negative predictive values are above 90% regardless of the location of the variant on the gene sequence, and even reaching 99.9% deep in introns. The lowest NPV is obtained for the polypyrimidine tract with 93.1%. It means that a negative prediction is highly reliable and that the variant should not be considered as spliceogenic.

SPiP has been designed for research but also diagnostic purposes. From a diagnostic point of view, sensitivity is key in order not to miss a pathogenic variant to be used for patient's care. Depending on the probability of splice alteration and the associated PPV of the altered motif(s), the user has all necessary information to decide whether a variant should be considered spliceogenic or not. Importantly, spliceogenic doesn't necessarily mean pathogenic, despite the 98.74% sensitivity obtained from the Clinvar data. Neither the level of alteration, *i.e*. total or partial effect, nor the exact alterations are predicted by SPiP. This is why positive predictions should be characterized at the RNA level for pathogenicity assessment before issuing any diagnostic report. Alternatively, negative predictions come with high confidence and prediction-based classification towards class 2 should be considered for negative variants.

Therefore, SPiP has the potential of a widely used decision-making tool to guide geneticists toward relevant spliceogenic variants in the deluge of high-throughput sequencing data.

MATERIAL AND METHODS

1. Evaluation set

This curated set was made of 2,784 variants from 213 genes with their RNA *in vitro* studies (2,402 published and 382 unpublished Supplementary Figure S3).

The published data were obtained from: (i) reviews of variants with their RNA *in vitro* studies (n = 943) [7], [21]–[23], (ii) the DBASS5 and DBASS3 databases (n = 110) [24], [25], (iii) the prospective collection of literature data by ENIGMA consortium [26] (n = 675) and (iv) by institute Cochin (n = 674). Each data was manually curated for genomic variant and for splicing alteration in main article. The 382 unpublished variants were studied for diagnosis purpose and collected by a collaborative effort of: i) the French splicing network of Unicancer Genetic Group (UGG) (n = 167), ii) the institute Cochin (n = 110), iii) the Inserm U1078 (n = 63), iv) the Inserm U1245 (n = 22), v) the laboratory of genetic of the hospital Saint-Louis-Lariboisière-Fernand Widal (n = 17) and vi) the Center of genomics of the university of Copenhagen (n = 3). RNA was extracted from whole blood collected on PaxgeneTM and/or lyphoblastoïd cell lines teated or untreated with the NMD inhibitor puromycin. Transcript analyses were based on RT PCR followed by Sanger sequencing and/or minigene assays [5], [27]–[31].

For statistical purposes, 4 classes of splicing defects were defined (Supplementary Figure S2): i) exon skipping gathered unique and multiple exon cassette skipping ii) intronic exonisation or pseudo exons

iii) intronic retention and iv) splice site shifting, for which we reported the splice site type (donor/acceptor) and the relative distance to the natural splice site. We also differenced the two main mechanisms of splice site shift: activation of non-natural splice site by the variant and alteration of natural splice site leading to the use of non-natural splice site. Quantitative aspects were not considered and both partial and total transcript alterations were considered as a unique splicing alteration.

2. Selection of bioinformatics tools

Among the vast number of available in silico tools, our selection criteria were based on i) availability at the time of this work, ii) possibility to implement in a pipeline (*i.e.* exclusion of tools only-online accessible) and iii) possibility of high throughput analyses (compatible with 'batch' analysis, reasonable runtime). Selected tools and splice motifs addressed are shown Supplementary Table S1.

a. Consensus splice site

As distinct 5'/3' prediction tools are based on distinct consensus motif lengths, we defined the smallest overlap consensus length to ensure homogenous comparison: -3; +6 for donor motif length and -12; +2 for acceptor motif length. The selection of tool dedicated to 5'/3'ss was performed on the well-characterized validation set described in Leman *et al.* [7] (n = 253 variants in 11 genes). We compared Splice Site Finder (SSF) [32], MaxEntScan (MES) [33], Human Splicing Finder (HSF) v3.0 [34], ADAboost/RandomForest [6], splicing-based analysis of variants (SPANR) [14] and Splicing Prediction in Consensus Element (SPiCE) [7]. These tools give quantitative values and we calculated the score variation between wild-type and mutated sequences, ADAboost/RandomForest, SPANR and SPiCE excepted, as they already computed a variation score. Then, scores were compared by ROC analysis (library 'ROCR') with R software.

b. Polypyrimidine tract (PPT)

The PPT was defined between the 13th and the 17th nucleotide upstream the 3' splice site. The reason is that 3' splice site predictors extend up to 12 nt in the intron and the branch point area was described as starting from 18th nt in intron [35]. To address the variant occurring in this region, we compared MES and SPANR, taking into account up to -20 and -300 nt in intron respectively.

MES score was used according the guidelines emitted by the Unicancer Genetic Group in 2012 [5] on variants occurring in consensus splice sites. Indeed, a decision-making threshold, corresponding to delta score at -15% was proposed to predict splicing alteration

c. Branch point

The optimal tool for branch point prediction was defined on a set of variants with their RNA *in vitro* studies occurring in the branch point area (-18, -44) [35]. This data collection was performed in the

scope of an in review benchmarking [8] of 5 BPs-dedicated tools: SVM-BPfinder [36], Branch Point Prediction (BPP) [37], Branchpointer [38], LaBranchoR [39] and RNA Branch Point Selection (RNABPS) [40]. Briefly, for this benchmarking on physio-pathologic data, we assayed two strategies to predict variant-alteration branch point: is variant located in the predicted branch point motif and is variant decreasing the score of predicted branch point. With the strategy optimal performances, we compared the bioinformatic tools. The details of this comparison are available at: https://dx.doi.org/10.21203/rs.2.12748/v1.

d. Exonic Splicing Regulators (ESR)

The optimal tool for the ESRseq, was determined from previous team work [9] that compared ΔtESRseq, ΔHZ_{EI} and SPANR on 22 variants occurring in exon 10 of *MLH1* gene. Due to this small set of variant, with completed this study by a recent published benchmark performed by [11]. This work had shown ΔtESRseq with the optimal performance on a set of 20 variants occurring in *BRCA1*, *BRCA2*, *NF1* and *DMD* genes among EX-SKIP, ΔtESRseq and ΔHZ_{EI}. Thus ΔtESRseq was used to score ESRseq. However, the initial publication of ΔtESRseq [41] did not preconized a decision-making threshold. An average decision threshold was estimated on our variants occurring in the range of scalable ESRseq [42] defined as the 120 exonic nt bordering start and end of exon. We excluded variants altering splicing other than exon skipping such as splice site shifting. Cross-validation was processed 1,000 times with a random sample of 100 variants used to set the optimal threshold (corresponding to maximal accuracy) and tested of the remaining variants.

e. Splice site creation or reinforcement (de novo/cryptic 5'ss/3'ss)

For the variant creating new splice site, we developed a new model to detect the use of cryptic splice site whatever the position of variant. This new model was a metascore based on logistic regression gathered the scores of consensus splice site and enhancers/silencers motifs. MES and position weight matrix, such as SSF, was selected for consensus splice site scores. Indeed, this combination was determined as optimal to consensus splice site [6], [7]. Enhancers/silencers motifs were scored by the ESR scores obtained by large scale minigene-spliced assay [42]. We trained and validated it on a set splice sites from the transcripts described in Ensembl data (download June 28th 2018) (n = 530,931 spl ice sites). As control data, we took all AG and GT motifs in these transcripts, corresponding to a comprehensive list of control AG/GT (n = 202,458,725 controls AG/GT). Thereby, this set had a ratio of true splice sites at 1:381. Two third of this data collection was used to train the model. The one third remaining of this collection was use as validation set. On this validation set we compared performances of this model to each individual score by ROC analysis. This step was also used to define the optimal decision-making threshold. The strategy to detect a splice site activation from this model was illustrate in Figure 3. Briefly, we compare the scores of potential splice site around the mutation between wild-type and mutated sequences. From this comparison, the tool considers only the splice sites with a

reinforcement of score or a new score apparition to the detection of splice sites. Then on these last splice sites, we applied the optimal threshold previously defined. Whether a splice site has a score above the threshold, we consider it as an activation of new splice site.

3. SPiP workflow

SPiP locates the variant within the genomic sequence to identify the relevant splice motifs and select the associated prediction tool(s). SPiP reports the altered splice motif with a probability of alteration according to the result of the relevant prediction tool. The transcripts database used by SPiP is the RefSeq database with the assembly genome version hg19 and hg38. The input of SPiP is the transcript ID (RefSeq) and the HGVS (Human Genome Variation Society) mutation nomenclature, ":"-separated, (e.g.: NM_007294:c.4096+3A>G). SPiP was developed to suit the Variant Call Format (VCF) v4.0 or later or standardized text file format as input. SPiP runs a R script to calculate the scores (available at https://sourceforge.net/projects/splicing-prediction-pipeline/ in a standalone version thus not supplemental installation is necessary to use SPiP. SPiP is available through Windows and Linux. The Windows version offers a friendly and easy to use interface. The Linux version allows high throughput analyses through parallel processing of a great number of variants (runtime is ~0.5 sec/variant/CPU on Linux CentOS (Supplementary Figure S9).

4. Estimation of splicing probability

In this part we estimated the risk that variants impact splicing according the prediction of SPiP. This estimation implies first the calculation of the probability that a variant impacts splicing. Several works have already estimated the proportion of spliceogenic variant, especially for exonic variants. The average proportion of variants affecting splicing in exon is approximatively 23.5 % with a range to 10.3% – 48.0 % (Supplementary Table S2). However, for intronic variants, the estimation of proportion of spliceogenic variants is tricky due to the rarity of these events and the publication bias. Indeed, the intronic regions are particularly biased in the rapported spliceogenic variants. Thus we used Bayes theorem to estimate the proportion of spliceogenic variants from our variants collection (see evaluation set). The canonical variants (+1,+2/-1,-2) were excluded due to their heavy impact on splicing. We splitted up variants in intronic variants and exonic variants. We controlled that the proportion of exonic spliceogenic variants was concordant with the literature (Supplementary Table S2).

The distribution of mutation between intronic and exonic variant was estimated from the 1000 genome data. The generic Bayes equation, with R the event variant in exonic region and S the event splicing alteration is:

$$P(S|R) = \frac{P(R|S) * P(S)}{P(R)}$$

Where: $P(S|R) = {}^{N\ spliceogenic\ variants\ in\ R}/_{N\ variant\ in\ R}$

$$P(R|S) = {}^{N\ spliceogenic\ variants\ in\ R}/_{N\ spliceogenic\ variant}$$

The $P(R|S)$ term is the probability that a spliceogenic variant occurs in exon and $P(S|R)$ corresponds to proportion of spliceogenic variant among exonic variant. These two terms were estimated from our collection of variant. We could define the equation to get the probability that variants alter splicing is:

$$P(S) = P(R) * \frac{N\ spliceogenic\ variants}{N\ variant\ in\ R}$$

For each estimated probability, we calculated the confident interval at 95 % according Wald method. To re-establish this probability in our dataset, we set as control data the frequent variant (Minor Allele Frequency > 5%) occurring in genes described in the evaluation set. The evaluation set gathering variants studied for diagnosis purpose therefore the genes described in this set were involved in human diseases. We hypothesized that frequent (>5%) variant cannot alter the functionality of these genes and so have no impact on splicing. The evaluation set plus control data permitted to estimate the probability of splicing alteration according to the SPiP prediction and the position in transcript.

5. Evaluation of variant pathogenicity by SPiP

From the ClinVar dataset, we selected the variants reviewed by an expert panel. Variants of unknown significance were excluded. Among pathogenic and likely pathogenic variants, truncating and missense variants were removed as their protein impact could explain their pathogenicity. The remaining variants were used for SPiP assessment.

2. Supplementary information

Supplementary Table S1: Bioinformatics tools studied in this work

Tool	Consensus 5'/3'ss	Polypyrimidine tract	Branch point	ESRs
SSF, SpliceSite Finder	+	+ (up to -14)	-	-
MES, MaxEntScan	+	+ (up to -20)	-	-
HSF, Human Splicing Finder	+	+ (up to -14)	-	+
SVM-BPfinder	-	-	+	-
EX-SKIP	-	-	-	+
ΔtESRseq	-	-	-	+
ADAboost/ RandomForest	+	+ (up to -12)	-	-
ΔHZ_{EI} (HEXplorer)	-	-	-	+
SPANR	+	+	-	+
BPP, Branch point Prediction	-	-	+	-
Branchpointer	-	-	+	-
SPiCE	+	+ (up to -12)	-	-
LaBranchoR	-	-	+	-
RNABPS	-	-	+	-

Supplementary Table S2: Proportion of exonic variants impacting splicing reported by the literature.

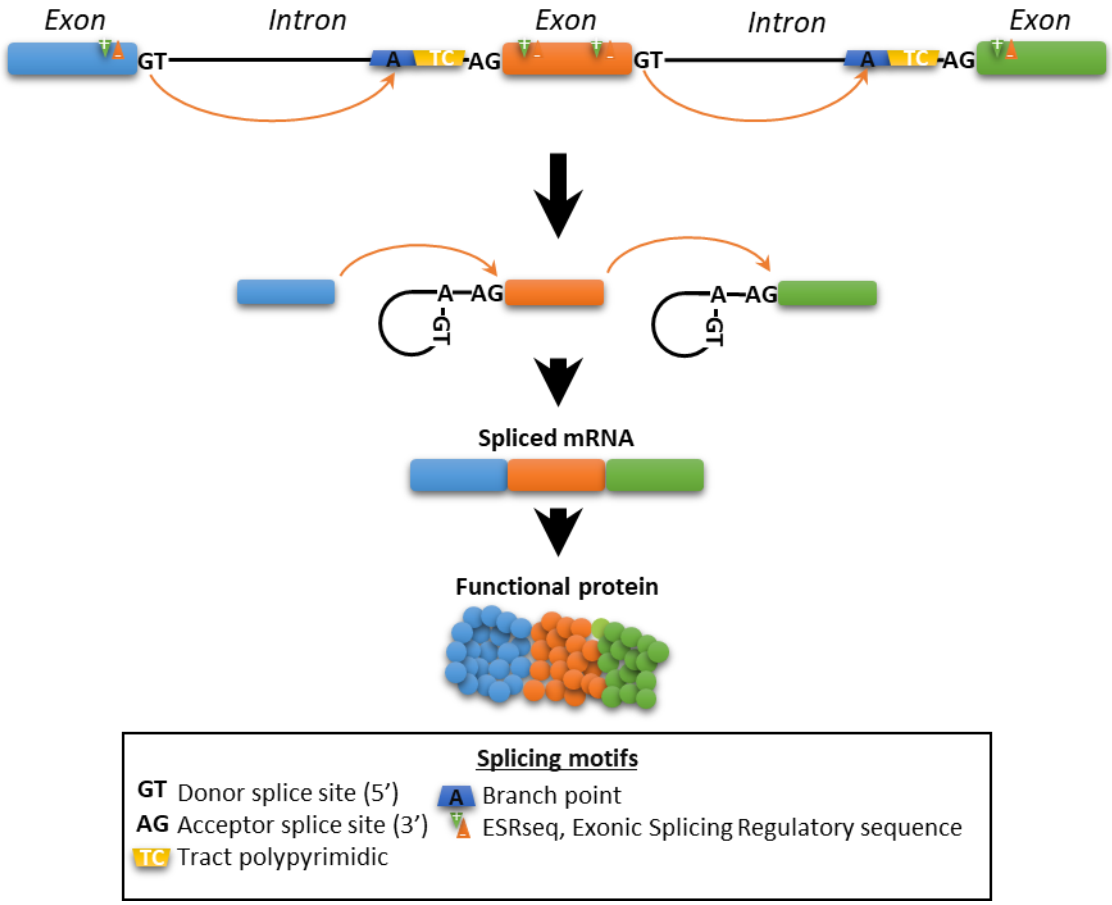
Number of spliceogenic variants/all tested variants	%	Method used	Reference
30/62	48.0%	Study of ATM variant by RT-PCR of LCLs (Exon+Consensus)	[43]
13/67	19.4 %	Spliced-minigene assays (*MLH1* and *MSH2*)	[44]
-/-	22.0 %	Simulated mutation from HGMD and intraallelic L1 Distance	[45]
7154/27681	25.8 %	Prediction of loss ESE/gain ESS of 27,681 HGMD mutations with 83 spliced minigene assays for validation	[46]
32/138	23.2 %	Spliced-minigene assays (*SMN1*)	[47]
513/4964	10.3 %	Massively parallel splicing assay (MaPSy)	[48]

Supplementary Table S3: Optimal strategy to predict branch point (n = 120 variants). (Motif 4-mers: TRAY). TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

	Is the variant decreased the score of predicted branch points? (best tool: SVM-BPfinder)	Is the variant located in motif 4-mer (TRAY) of predicted branch points? (best tool: BPP)
TP	29	32
FP	13	7
TN	69	75
FN	9	6
Accuracy	81.67 %	89.17 %
Sensitivity	76.32 %	84.21 %
Specificity	84.15 %	91.46 %

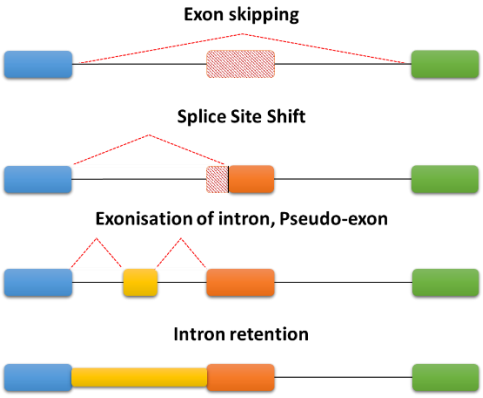
Supplementary Table S4: Overall performances of SPiP on 2,936 variants reported by ClinVar. TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

	SPiP prediction
TP	313
FP	596
TN	2023
FN	4
Accuracy	79.56%
Sensitivity	98.74 %
Specificity	77.24 %

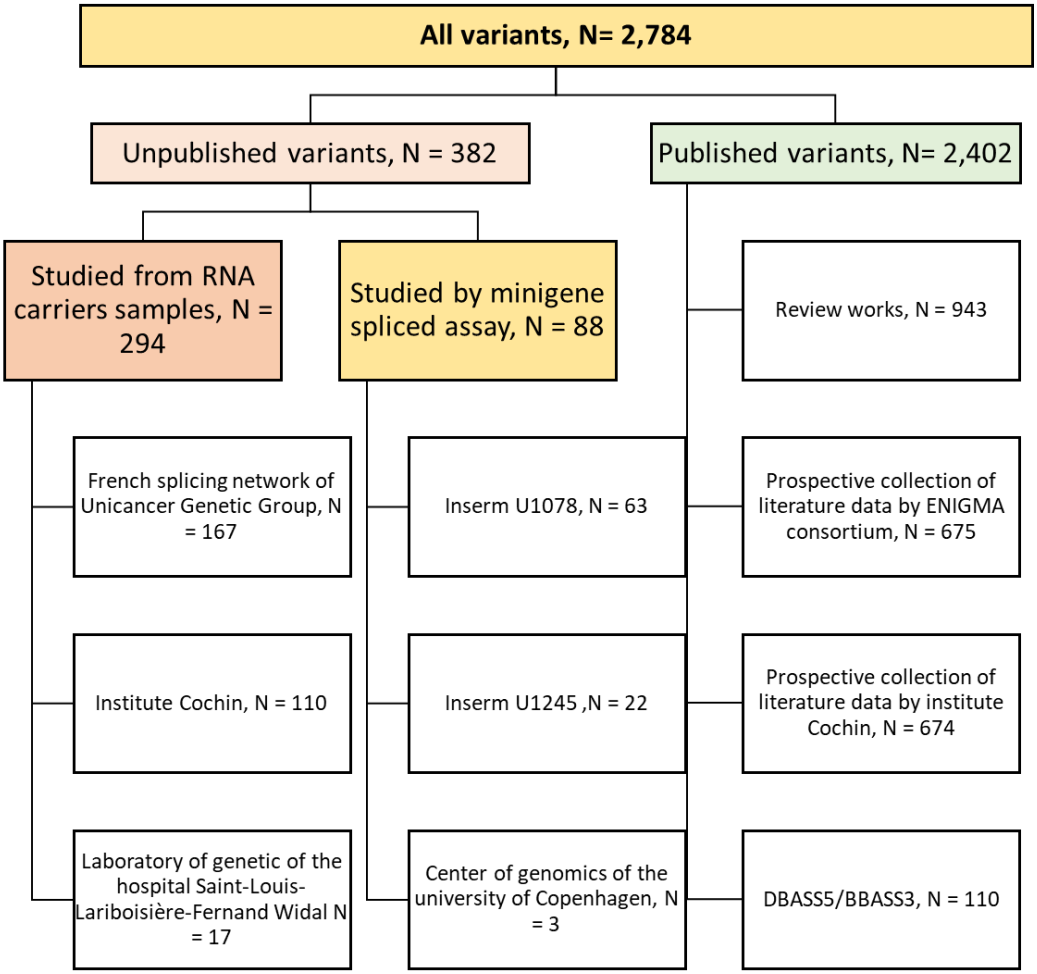


Supplementary Figure S1: Splicing mechanism and motifs.

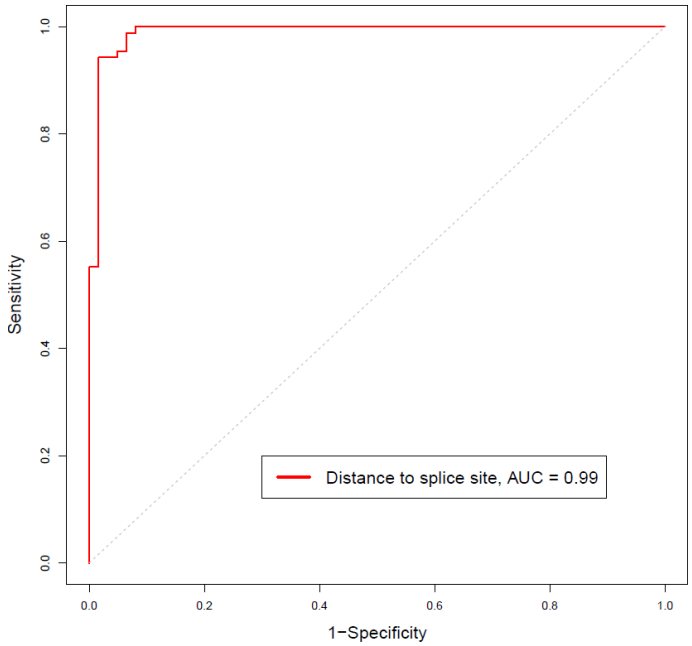
The donor splice site defines the exon/intron junction, with two highly conserved nucleotides (GT). The acceptor splice site delineates the intron/exon junction, with a highly conserved dinucleotide (AG). The branch site is a short motif upstream the 3'ss that includes the branch point (BP) adenosine. These BPs are mainly located in area between -44 and -18 nt of the natural 3'ss [35]. Separating the 3'ss and the BPs area, there is a cytosine and thymidine rich sequence called polypyrimidine tract (PPT). The identification of these motifs depends also of short motifs (6-8 nt) defined as splicing regulatory elements (SREs). Briefly, these motifs are binding signals recognized by RNA-binding proteins, mostly SR (serine and arginine rich) proteins. The SREs can be enhancers or silencers for the identification of splice sites by the spliceosome. The SREs act mostly in exonic region and in this region are called exonic splicing regulatory sequences (ESRseq) [42].



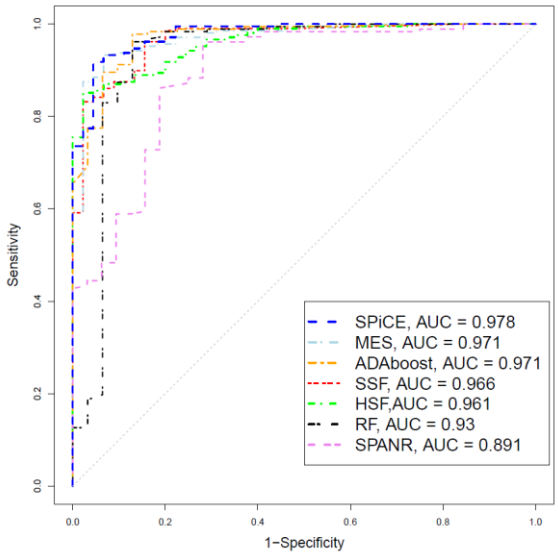
Supplementary Figure S2: Illustration of splicing alteration induced by a variant



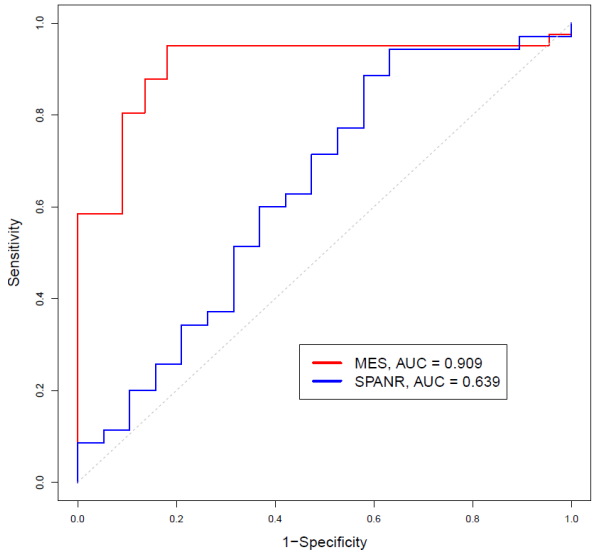
Supplementary Figure S3: Collection of 2,784 variants with RNA *in vitro* studies.



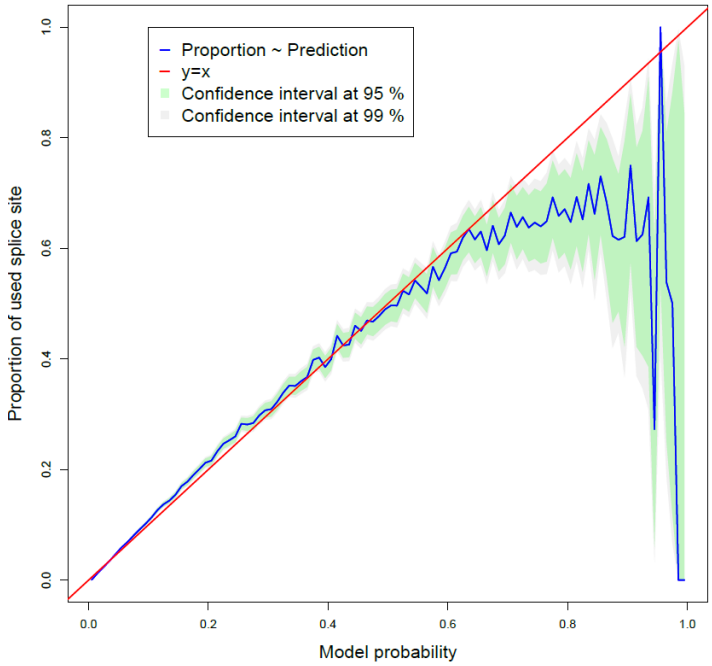
Supplementary Figure S4: Distance between intronic variants and splice sites to discriminate the splice site shift to pseudo-exon use.



Supplementary Figure S5: Comparison of bioinformatics tools 3'/5'ss dedicated on 253 variants.

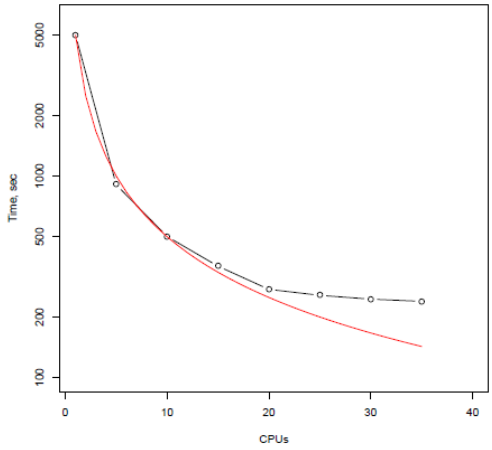


Supplementary Figure S6: Comparison of MES and SPANR scores on the polypyrimidine tract variants, N = 63 variants



Supplementary Figure S7: Correlation between the model probability and the proportion of used splice sites on the validation data (N = 67,663,176 splice sites)

Figure S8: available in pdf file FigS8_VPP_position.pdf



Supplementary Figure S9: Assesment of runtime of SPiP on 10,000 variants with different number of CPUs from Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz, 40 CPUs and 258 Go of RAM.

IV. ANNEXE D SUPPLEMENTARY INFORMATION: SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from target RNAseq data.

Fig. S1: The bioinformatics pipeline that computes the read count matrix from fastq files.
The shell script of this pipeline, as well as an example of fastq files, exons coordinates BED files, perl scripts and installation instruction are available at https://github.com/raphaelleman/SpliceLauncher.

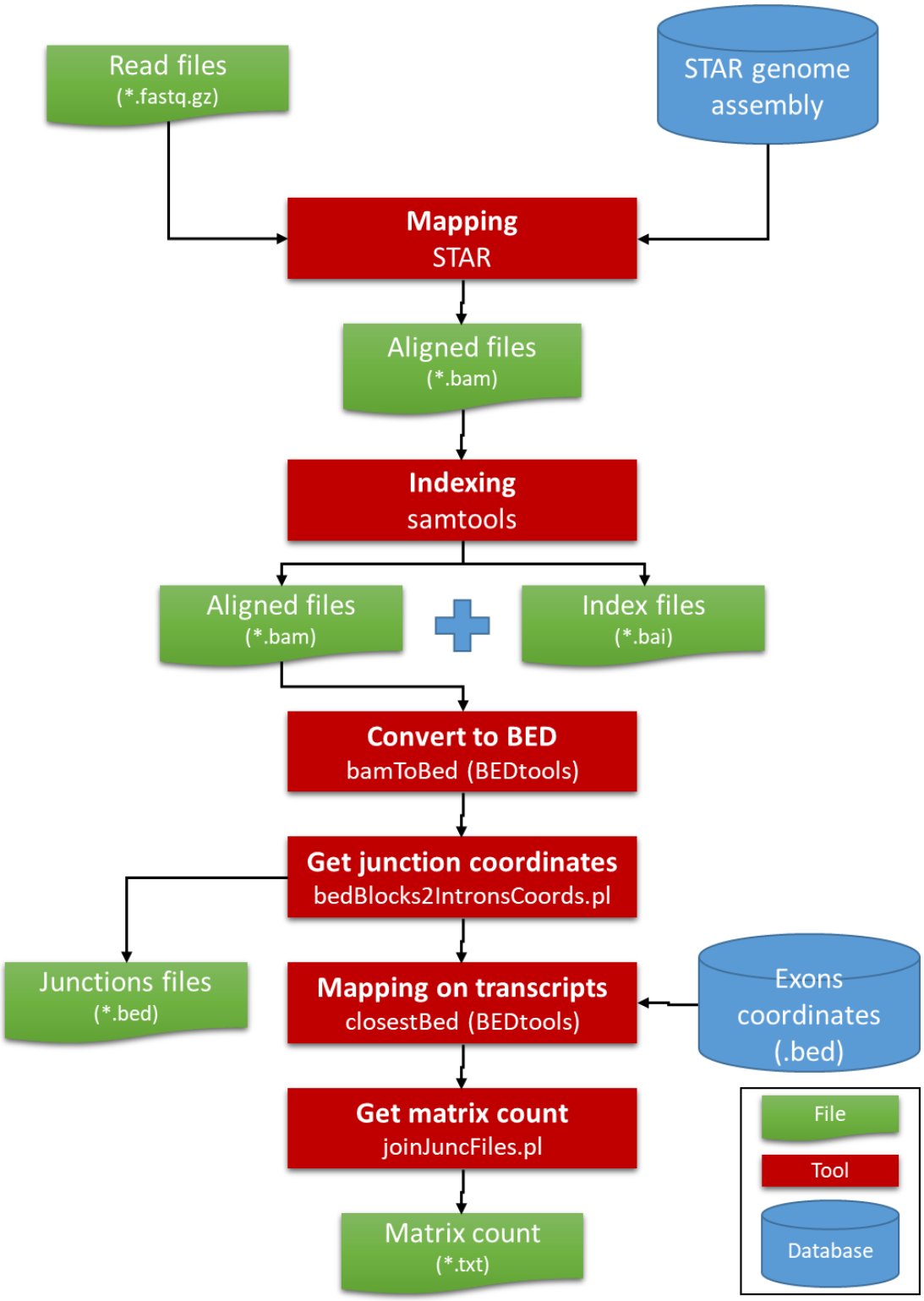


Fig. S2: The junction names are attributed according to the reference transcript and adapted from the recommendation of [49].

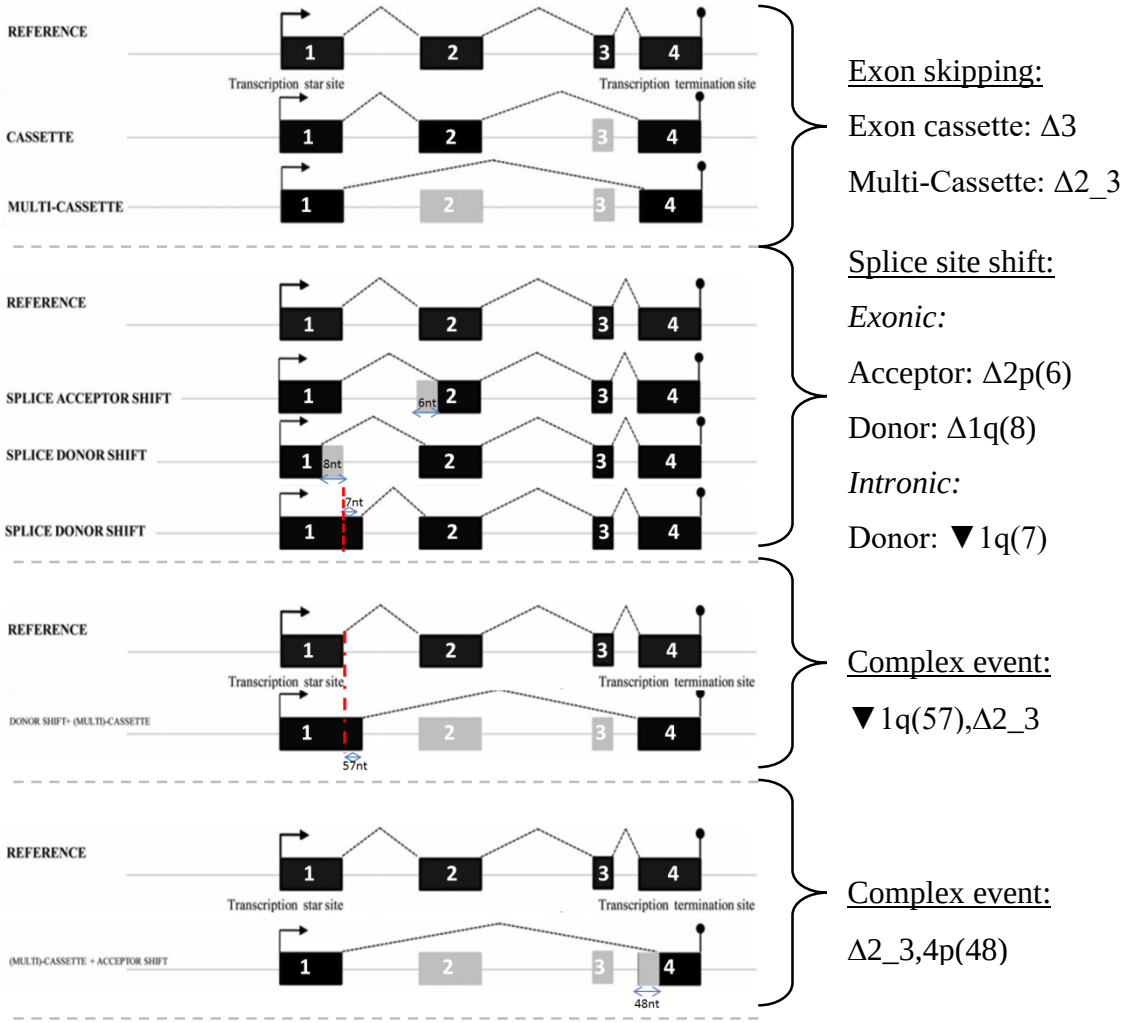


Fig. S3: The calculation application for each type of alternative splicing events [50].

If the alternative transcripts have bi-allelic expression, the relative expression moves towards +∞, if the alternative transcripts have mono-allelic expression, the relative expression moved towards 100%. Exons are represented by black boxes, constitutive junctions are represented by green lines and alternative junctions are represented by red lines. **(a)** cryptic exon (CE) inclusion; **(b)** exon skipping (E_{n-1}); **(c)** multiple exon skipping (E_{n-1} E_{n-2}); **(d)** splice intronic donor shift (I_a); **(e)** splice intronic acceptor shift (I_b); **(f)** splice exonic donor shift (E_{xa}) and **(g)** splice exonic acceptor shift (E_{nb}).

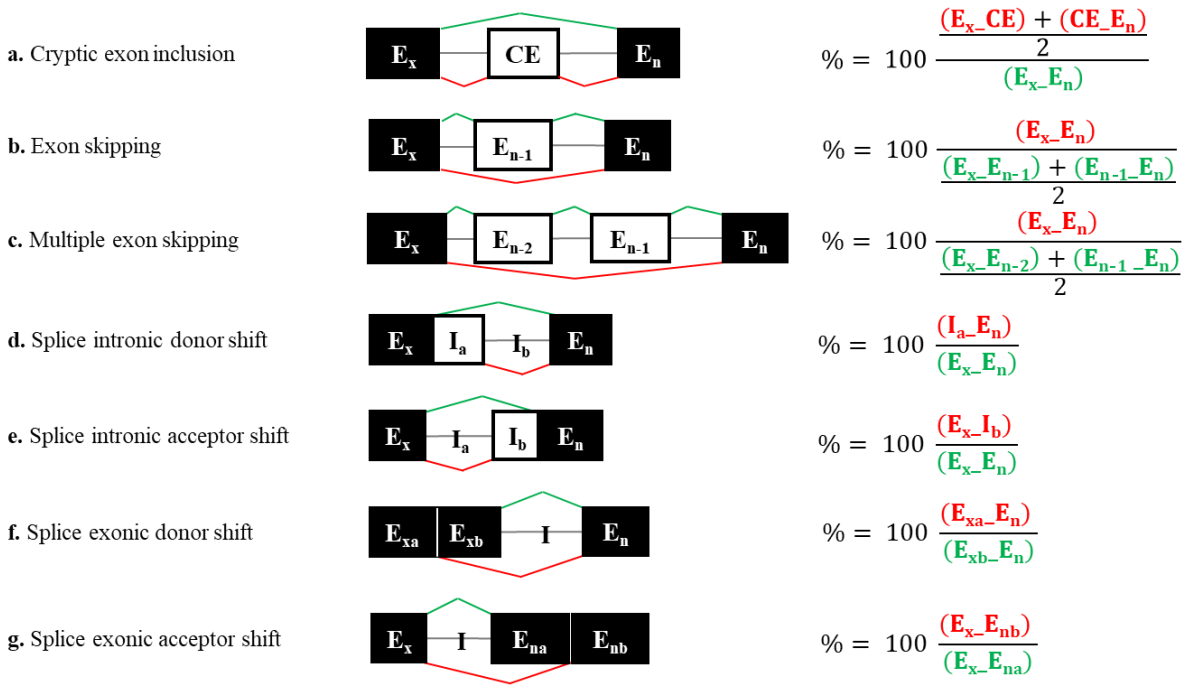


Fig. S4: Example of a pdf report generated by SpliceLauncher for the *BRCA2* gene.

Only the genes having at least 100 reads mapping on its physiological junctions and constituted of more than 2 exons are shown. Only the junctions above the threshold (here 1 %) are represented. If the statistical analysis is performed, the junctions found significantly expressed are show in bold and italic as follow: ***Δ4 : 63.82 % ******, with *: p-value 0.05-0.01; **: p-value 0.01-0.001; ***: p-value <0.001.

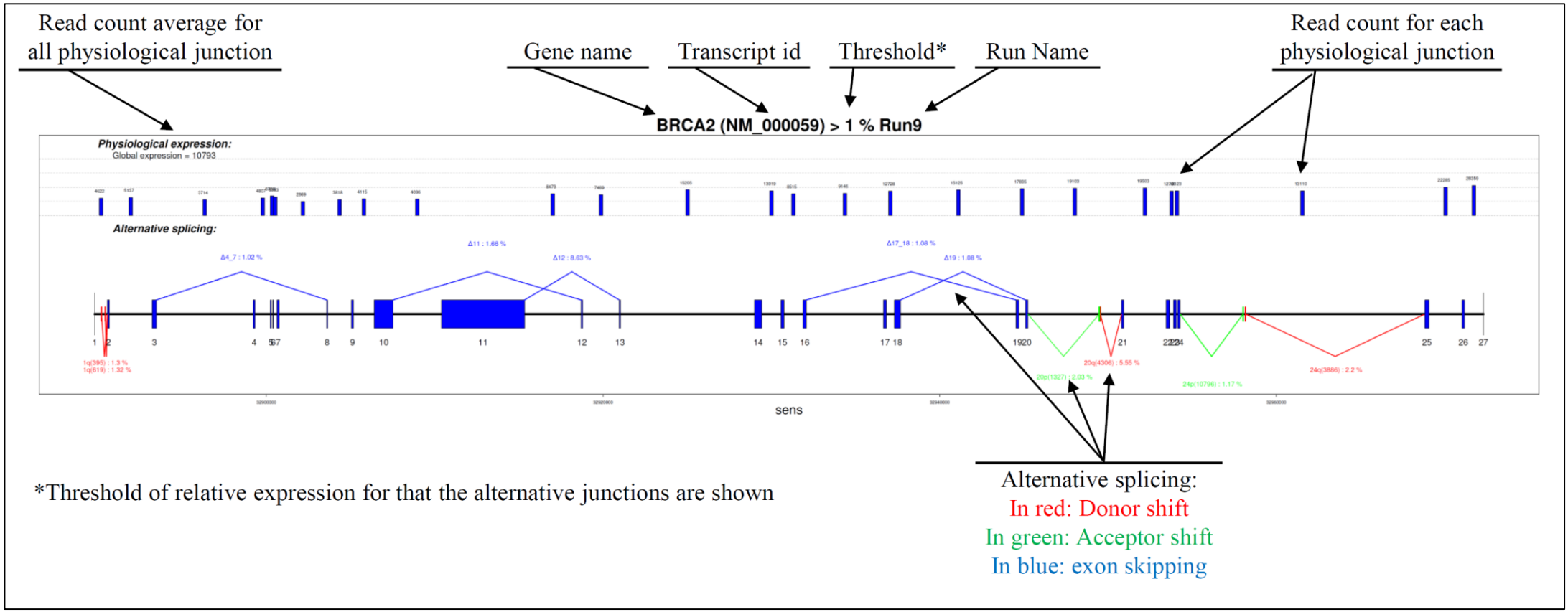


Fig. S5: Screenshot of the visualization of a junction alignment in the UCSC genome browser.
In Blue are represented the exon skipping junctions, in red the donor shift junctions and in green, the acceptor shift junctions.

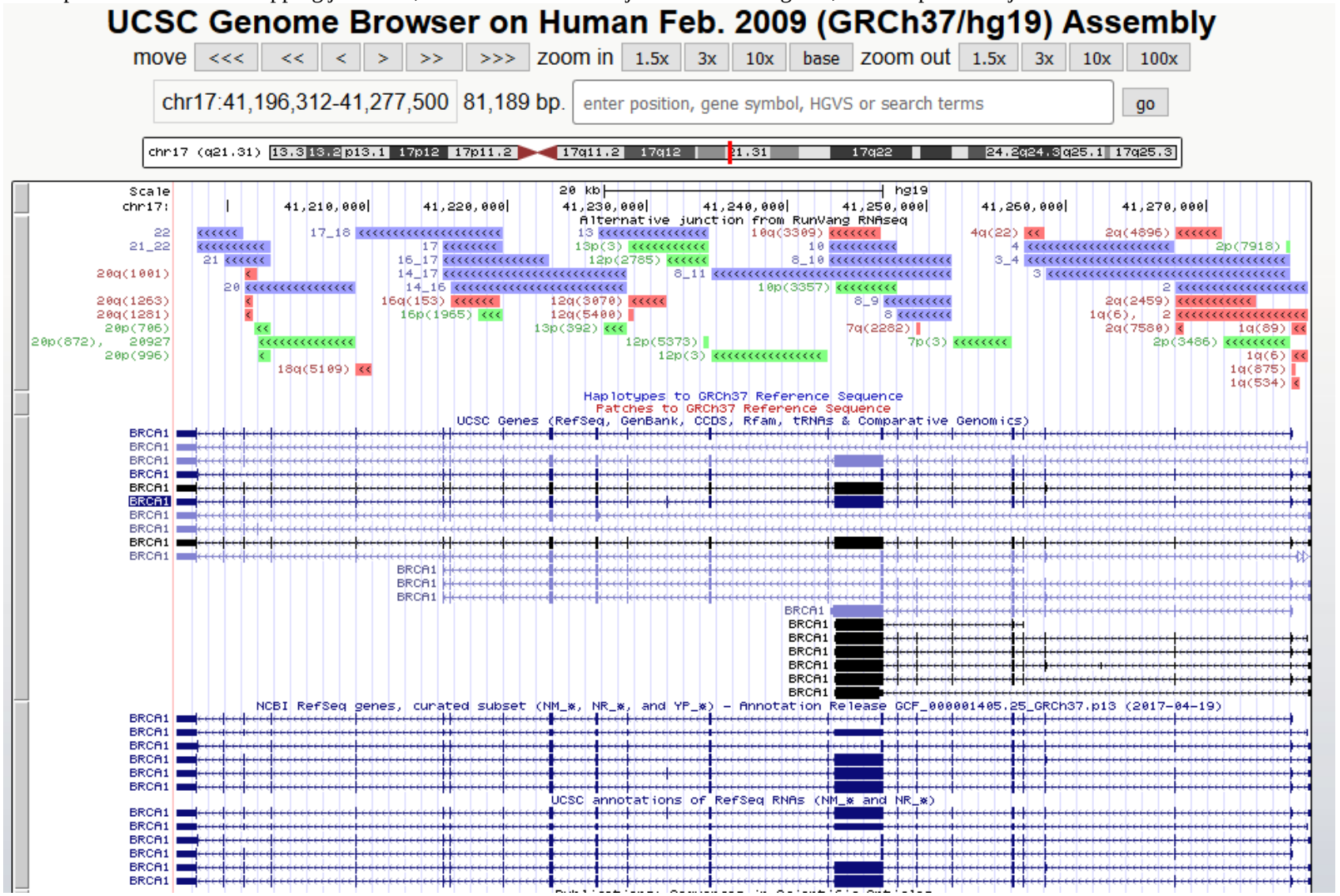


Fig. S6: Example of SpliceLauncher command lines

```
$ ## Example of RNAseq pipeline + Splicelauncher analysis
$ ## Used software:
$ # STAR v2.6
$ # BEDtools v2.24
$ # SAMtools v1.9
$ ## Annotation files:
$ # Genome (hg19) download (08/22/2018) at
ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh37_latest
/refseq_identifiers/GRCh37_latest_genomic.fna.gz
$ # RefSeq annotations (hg19) download (02/15/2019) at
ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh37_latest/refseq
_identifiers/GRCh37_latest_genomic.gff.gz
$ ls ./dataTest/fastq/*.fastq.gz # FastQ files used by SpliceLauncher
      A7BDE-785-01-07-111-0075_T.fastq.gz
      A7BDE-785-02-07-111-0075_U.fastq.gz
      A7BDE-785-03-07-111-0082_T.fastq.gz
      A7BDE-785-04-07-111-0082_U.fastq.gz
      A7BDE-785-05-07-111-0085_T.fastq.gz
$ # RNAseq pipeline + SpliceLauncher analysis
$ bash ./ SpliceLauncher.sh --runMode Align,Count,SpliceLauncher \
 -F ./dataTest/fastq/ \
 -O ./testSpliceLauncher/ \
```

Fig. S7: The mutation c.4096+3A>T in *BRCA1*. A. RT-PCR Long-Range exon 10-12 was used due to the amplicon size > 3.2 kb (primer F: GAT GAA ATC AGT TTG GAT TCT G, primer R: TGT CAC TCT GAG AGG ATA GC). The reverse transcription was performed on 2 µg of total RNA by the MuLV Reverse Transcriptase (#N8080118 Life Technology). The reaction mix (qs 40 µL) contained: 2µL of MuLV RT (50U/µL), 8µL of $MgCl_2$ (25mM) and 4µL of Buffer II (10x) (#N8080130 Life Technology), 2µL of random hexamers (50µM, #N8080127 Life Technology), 16µL of dNTP (10mM, #430344 Life Technology), 2µL of RNase inhibitor (2000U, #N8080119 Life Technology). The mix was incubated 10 min at 25°C, 60 min at 45°C and 10 min at 70°C. The PCR long-range used the Long Range DNA polymerase kit (#BR0300301 Biotechrabbit) according the manufactory's instructions on 2 µL of cDNA. Cycling conditions were as follows: 95°C for 2 min and 35 cycles of 95°C for 30 s, 55°C for 45 s and 68°C for 4 min and final elongation 72°C for 5 min. C: Controls samples, P: c.4096+3A>T carriers, +/-: with/without puromycin, inhibitor of nonsense mediated decay. Each band was identified by Sanger sequencing. B. Result of RNAseq analysis by SpliceLauncher, with two biological replicates of c.4096+3A>T and eight controls, with the p-values.

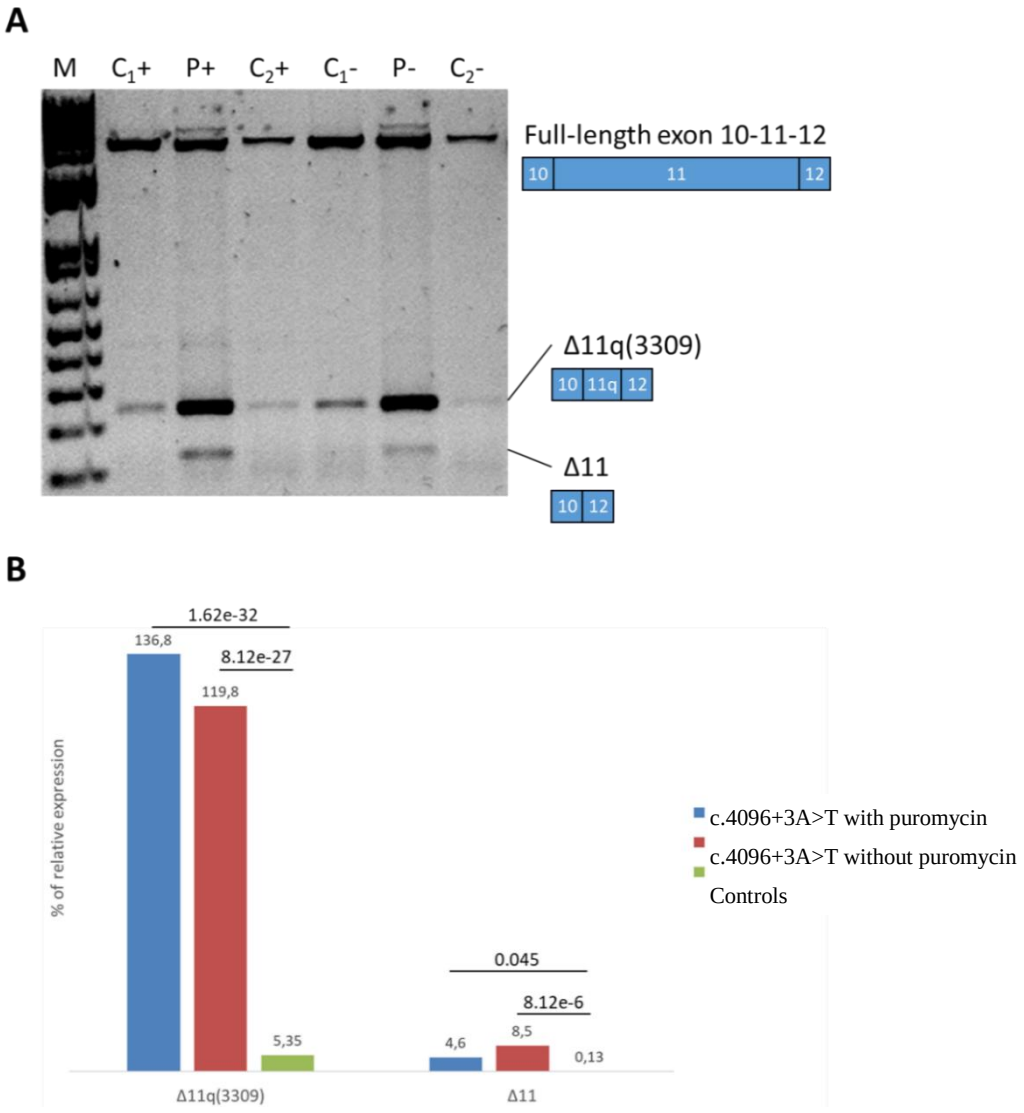


Table S1: The scheme of annotation file used by SpliceLauncher to align the junction on transcripts.

Column name	Example	Description
Gene	BRCA2	Gene symbole
Strand	+	Strand of transcript
gCDSstart	32890597	start of coding sequence
gCDSend	32972907	end of coding sequence
transcript	NM_000059	Name of the transcript
Chr	chr13	Chromosome name
idEx	1	exon number
lenEx	188	exon length
gStart	32889616	exon start in genomic coordinates
gEnd	32889804	exon end in genomic coordinates
cStart	-227	exon start in transcriptomic coordinates
cEnd	-40	exon end in transcriptomic coordinates

Table S2: The scheme of results saved by SpliceLauncher in excel format or in tabulated-text format (optional).

Column name	Example	Description
Conca	chr19_58858391_58858718	"_"-separated chromosome name, start position and end position
chr	chr19	Chromosome name
start	58858391	Start position (end position for '-' strand item)
end	58858718	End position (start position for '-' strand item)
strand	-	Strand of the junction
Strand_transcript	reverse	Strand of the transcript
NM	NM_130786	Name of the transcript
Gene	A1BG	Name of the gene
XXX … XXX	0 … 1	Matrix of read count
P_XXX … P_XXX	0 …0.2	Matrix of relative expression (%)
Event_type	3AS	Letter code: "SkipEx" exon skipping; "3AS" 3' cryptic ss; "5AS" 5' cryptic ss; "Physio" physiological junction; "NoData" no annotated junction
AnnotJuncs	Δ8p(4)	Name of junction*
cStart	c.1480	Start transcriptomic position
cEnd	c.1485	End transcriptomic position
mean_percent	0.438139676984821	Average of relative expression for each junction
read_mean	0.555555555555556	Average of read count for each junction
nbSamp	1	Number times of junctions is observed
DistribAjust	Gamma	Distribution law of junction (if statistics analysis)
Significative	NO	Junction contains outliers (YES/NO) (if statistics analysis)

* Event too complex: junctions whose coordinates no matching with at least one known acceptor or donor splice site.
Outside transcript: junction with the start or the end position outside the end or the start of the transcript

Table S3: The BED format generated by SpliceLauncher.

By default, the track line of the BED files is: track name=<run name> type=bedDetail Description="Alternative junction from <run name> RNAseq" db=hg19 itemRgb="On" visibility="full". It is possible to select another genome assembly by changing the value of the db argument. Replacing itemRgb="On" by itemRgb="Off" will switch off junction colors, and using useScore=1 will color junctions with shade of gray corresponding to junction score.

Column name	Example	Description
chr	chr19	Chromosome name
start	58858391	Start position (end position for '-' strand item)
end	58858718	End position (start position for '-' strand item)
name	Δ8p(4)	End position (start position for '-' strand item)
score	0.438139676984821	Average of relative expression (%) for each junction
strand	-	Strand of the junction
thickStart	58858391	Start position (end position for '-' strand item)
thickEnd	58858718	End position (start position for '-' strand item)
itemRgb	80,255,80	An RGB value of the form R,G,B
blockCount	1	Number of junctions, set to 1
Annot1	c.1480_ c.1485	Transcriptomic coordinates
Annot2	Sample1: 0.2|Sample2: 0.6|…	Relative expression (%) of the junction for each sample

Table S4: Summary of Chaligné *et al*., (GSE62966) results data [51].

The bioinformatics pipeline to compute the read count matrix took 8 h (1.6h/sample) on Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz, 40 CPUs and 258 Go of RAM. A total of 903,345 junctions were identified. SpliceLauncher took 6h including 4 h to annotate junctions and to calculate the relative expression and 2 h to display the graphical representation of alternative junctions. Among the 411,285 junctions, 75.4 % were annotated (310,316/411,285). From 34,227 transcripts, SpliceLauncher selected 21,388 transcripts as having one transcript by gene.

Junction types	Number	Read count (average [min; max])
Physiological	145,834	94.2 [0.2; 29617.8]
Exon skipping	35,151	4.9 [0.2; 9000]
Donor shift	61,656	6.2 [0.2; 7415.2]
Acceptor shift	67,675	4.9 [0.2; 8997.3]
Too complex event	36,650	3.9 [0.2; 2492.2]
Outside transcript	64,318	11.7 [0.2; 14565.8]

V. ANNEXE E SUPPLEMANTARY INFORMATION : Alternative Splicing and ACMG-AMP-2015 Based Classification of *PALB2* Genetic Variants: an ENIGMA Report

1. Supplementary methods

1. Identification of alternative splicing events

1.1 Targeted RNAseq experiments performed by Laboratory 1 (Clinical Biology and Oncology Laboratory, Cancer Center François Baclesse, Normandy University Caen, France): The overall RNAseq methodology has been described previously [50]. In brief, Agilent eArray (SureDesign; Agilent) was used to design SureSelect library baits targeting all reference exons in *PALB2* (and 10 other genes of interest). Libraries were sequenced on a NextSeq500 (Illumina, San Diego, USA) using the high-output paired-end 2x151 bps program, with 10 to 18 samples per run.

RNA was isolated from 73 unrelated samples (57 LCLs, 2 PaxGene tubes, 9 sLEU, 3 non-malignant breast tissues, and 2 HMEC cells. A total of 5 independent RNAseq experiments involving respectively 16, 18, 17, 10 and 12 samples were performed. The percentage of junction reads supporting alternative splicing events was calculated as previously described [50]. Overall, the analysis produced 14, 213, 304 reads aligned to exon-exon junctions in the *PALB2* locus. In total, up to 1326 putative splicing events were supported by at least one read. Yet, only events detected in four or more unrelated samples, detected in 3 or more independent RNAseq experiments, and representing on average > 0.00769% of junction reads were considered for further analyses. These (arbitrary) cutoffs were nonetheless based on previous experience characterizing *BRCA1* alternative splicing by targeted RNAseq and RT-PCR capillary electrophoresis [50], [52].

1.2 Whole Transcriptome RNAseq experiments performed by Laboratory 2 (Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Sweden): RNA was isolated from 9 normal breast tissue samples from unrelated women with breast cancer (one women tested for *PALB2* germ-line pathogenic variants, with negative results), and two normal fimbriae tissue samples from prophylactic oophorectomies performed in two likely post-menopausal women that have not had cancer. Methods for tissue sampling, preservation, RNA extraction, sequencing library preparation and sequencing of the breast tissue samples has been described previously [53]. Briefly, fresh breast tissue samples were preserved using RNAlater (Ambion). RNA was extracted using the AllPrep method (Qiagen). Sequencing libraries were prepared using a modified version of the dUTP method that preserves the strandedness of the RNA molecules in sequencing. Libraries were paired-end sequenced on an Illumina HiSeq 2000 (2x50bp, 2 samples) or NextSeq 500 sequencer (2x75bp, 7 samples). Two libraries were prepared from one of the breast tissue samples. The data from the two libraries were pooled in the bioinformatics analysis. The fimbriae tissue samples were fresh frozen. RNA was extracted using the AllPrep method (Qiagen). Sequencing libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit according to manufacturer's instructions (Illumina, San Diego, CA). Libraries were paired-end sequenced on an Illumina NextSeq 500 sequencer (2x75 bp).

Sequences from each library were trimmed and filtered to remove adapters and low quality bases using Trimmomatic and to remove reads that align to ribosomal RNA/DNA (GenBank loci NR_023363.1, NR_003285.2, NR_003286.2, NR_003287.2, X12811.1, U13369.1), phiX174 Illumina control (NC_001422.1), and sequences contained in the UCSC hg19 RepeatMasker track (downloaded 14 March 2011) using bowtie (with default parameters except -k 1 --phred33 --local).

Reads were aligned using STAR (version 2.4.1d) [54] to the human genome reference GRCh37. The genome index was created by using reference genome and annotation from Ensemble, distributed by iGenomes. The only adjusted parameter for index construction was sjdbOverhang 91 [54]. Other explicitly adjusted parameters used with STAR include outFilterMultimapNmax 2, outFilterMismatchNmax 20 and chimSegmentMin 0.

Overall, the analysis produced 5,267 reads aligned to exon-exon junctions in the *PALB2* locus (4,771 reads in breast and 496 reads in fimbriae samples). Up to 9 putative alternative splicing events were supported by at least one read, all them considered for further analyses.

1.3 Whole Transcriptome RNAseq experiments performed by Laboratory 3 (Department of Pathology and Biomedical Science. University of Otago Christchurch. New Zealand): Whole transcriptome sequencing analysis was undertaken on a LCL derived from a healthy female, cultured with and without cycloheximide treatment, as described previously [55]. cDNA libraries were constructed from total RNA using a TruSeq Stranded total RNA sample preparation kit (Illumina, San Diego, CA) following manufacturer's instructions, and sequenced on a HiSeq2000 (Illumina). Reads were mapped to the Homo_sapiens.GRCh37.72 reference genome, downloaded from Ensembl, using the two pass approach of the STAR (Spliced Transcripts Alignment to a Reference) aligner [54] using the default settings, except maximum intron length was set to 100,000. Detected splice junctions were extracted from STAR's SJ.out file for further analysis. Reads were mapped also using TopHat2 [56].

Overall, the analysis produced 9,663 reads aligned to exon-exon junctions in the *PALB2* locus (7,418 reads with cycloheximide and 2,245 reads without). Up to 20 putative alternative splicing events were supported by at least one read, all them considered for further analyses.

1.4 RT-PCR/CE experiments performed by Laboratory 4 (Molecular Oncology Laboratory. Academic Hospital San Carlos. Madrid. Spain): CE scanning was performed in RNAs extracted from up to 10 LCLs, 10 BLOOD samples, and commercially available breast and ovary RNAs. The methodology has been described previously [52], [55]. In brief, 300-700ng of total RNA were reverse transcribed with Prime-Script RT kit (TaKaRa Biotechnology, Japan) following manufacturer's protocol (the kit includes a mixture of random and OligodT primers). Multiple combinations of forward and reverse primers located at exonic regions defined by the reference transcript ENST00000261584 (**Supplemental Table 3**) were used to amplify cDNAs with FastStart polymerase by Roche. A PCR performed with a particular combination of primers will be referred throughout the text as a particular PALB2 splicing assay. The products generated by the various assays were analyzed by

conventional Br-Et agarose electrophoresis, and by capillary electrophoresis (CE). CE analyses were performed in an ABI 3130 genetic analyzer (Applied Biosystems), using LIZ-500/600 as internal size markers. Size calling was performed with GeneMapper v4.0 software (applied biosystems). Some events were captured by one assay, while others were captured by two or more overlapping assays (**Supplemental Table 3**). The RT-PCR/CE analysis involved a total of 1474 *data points* (one data point defined as each technical replica of an individual splicing event assayed in one sample). *RT-PCR/CE Coverage* (defined here as data points per splicing event tested) ranged from 8× to 80× (31× on average). *Detection Rate* (% of positive *data points*) ranged from 0% (splicing event not detected) to 94% (21% on average).

1.5 PALB2 whole-gene CloneSeq splicing analysis performed by Laboratory 5 (ATG Lab, Ambry Genetics, USA): For CloneSeq, blood was drawn in PAXgene Blood RNA Tubes and stored according to the manufacturer's recommendations (PreAnalytiX, Hombrechtikon, Switzerland). RNA was extracted using the PAXgene Blood RNA Kit according to the recommended protocol (PreAnalytiX). cDNA was generated using the SuperScript IV First-Strand Synthesis System (Thermo Fisher Scientific, Chino, CA, USA). The whole-gene CloneSeq splicing analysis methodology has been described recently ([57], in press, https://www.frontiersin.org/articles/10.3389/fonc.2018.00286/abstract). In brief, PCR was performed using either Platinum SuperFi PCR Master Mix (Thermo Fisher Scientific) or HotStarTaq Master Mix (Qiagen, Valencia, CA, USA). PCR products were cloned into pGEM-T Easy and transformed into bacteria according to the manufacturer's recommended protocol (Promega, Fitchburg, WI, USA). All colonies on a plate were scraped and suspended in PBS. Plasmids were extracted with the GeneJET Plasmid Miniprep kit (Thermo Fisher Scientific). CloneSeq libraries were constructed according to the protocol outlined by KAPA Biosystems (Wilmington, MA, USA) using the Hyper Prep kit. The size and concentration of the DNA library were determined using the TapeStation 2200. Massively Parallel Sequencing (MPS) was performed on an Illumina MiSeq, which generated 2×250 paired-end reads. Sequencing reads were aligned to the hg19 reference genome and analyzed using Ambry's Bioinformatic Pipeline. Abnormal transcripts levels were then measured as a "percent spliced in index" (PSI). PSI demonstrates the ratio between reads including or excluding exons, indicating how efficiently sequences of interest are spliced into transcripts [58]. The splicing events with "number of reads supporting alternative splicing event" < 20, or "number of all reads in the region covering splicing event" < 50, or "percent of splicing event" < 2.5% were filtered out. The main rationale behind these (arbitrary) cutoffs being that any transcript under these thresholds is very unlikely to be physiologically relevant. The same methodology was used to analyze splicing in carriers of c.2559G>C, c.3113+5G>C, c.3350+4A>C, and c.3350+5G>A variants **(see Supplemental Figure 6C)**.

2. Annotation of alternative splicing events

2.1 Descriptive annotation: We described all alternative splicing events and predicted protein products according to the HGVS guidelines (varnomen.hgvs.org), using the reference transcript Ensembl ENST00000261584.8 (NCBI reference sequences NM_024675.3 and NP_078951.2). For the sake of simplicity,

we also identified splicing events with a simplified code that combines the following symbols: Δ (skipping of exonic sequences present in the reference transcript), ▼(inclusion of intronic sequences not present in the reference transcript), p (proximal, 5' end of an exon, acceptor shift), and q(distal, 3' end of an exon, donor shift). When necessary, the exact number of nucleotides skipped (or included) is shown in brackets. See **Supplemental Figure 2** for further details.

2.2 Confidence of the finding: Alternative splicing events supported by a minimum of two laboratories, two independent samples, and two methodologies (RNAseq and/or RT-PCR/CE analyses and/or CloneSeq) were annotated as *high-confidence* naturally occurring alternative splicing events. Exceptionally, the exon 5 acceptor shift Δ(E5p24) was annotated as *high-confidence*, even though it has not been detected by RNAseq. Yet, we think that the quality of the RT-PCR/CE data supporting this finding is reassuring **(Supplemental Figure 3)**.

Other events were annotated as *lower-confidence* alternative splicing events (i.e. due to limited data supporting the findings, we cannot exclude technical artifacts and/or biological outliers). The list of 44 *lower-confidence* alternative splicing events detected in BLOOD includes two putative chimera read-through transcripts, one of them linking the penultimate exon of the upstream *loci* (*PALB2* exon 12) with the second exon of the downstream *loci* (NDUFAB1), a typical feature of read-through transcripts [59]8]. Many *lower-confidence events* have been detected by target RNAseq (N=41), but not confirmed by RT-PCR/CE (N=12), or not tested due to technical limitations of this approach (N=29). Others have been imputed from RT-PCR/CE data (N=3), but no direct confirmation by RNAseq has been obtained. Therefore, the amount of experimental evidences supporting lower confidence splicing events is variable (see Supplemental Table 2 for further details). Up to 19 lower confidence events were detected as well in BREAST, but only one was detected in OVARY (Supplemental Table 2).

2.3 Coding potential: The impact of each splicing event on the open reading frame (ORF) was analyzed with ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/). We performed annotations based on the following premises: (i) Functional domains spanning residues 9 to 200 (Coiled-coil and DNA binding), 395 to 561 (ChAM/ DNA binding) , 611 to 738 (MRG15), and 867 to 1186 (WD40-type Δ-propeller) are essential for PALB2 activity [60]–[62], (ii) The seven-bladed WD40-type Δ-propeller has the linear topology seen in most WD40-reapt domains (28 Δ-strands forming 7 four-stranded antiparallel Δ-sheets). This domain folds in a toroidal structure stabilized by interactions between the C-terminal WD40 Δ-strand 7C and the N-terminal Δ-strand 7D. This structural feature explains that PALB2 Y1183X is pathogenic, despite removing only the last three residues of Δ-strand 7C [63], [64]. Blades 4 and 5 are critical for PALB2-BRCA2 interaction [63]. (iii) Functional studies have identified critical PALB2 residues. In particular: coiled-coil residues L21, Y28, K30, L35 and R37 are critical for PALB2-BRCA1 interaction [65], [66], residue T413 is critical for PALB2 association with chromatin [67], WD40 Δ-propeller residues L939, T1030, and L1143 are relevant for stability and/or DNA double-strand break-induced Homologous recombination [68], WD40 Δ-propeller residue A1025, located at the bottom of the pocket formed by WD4 and WD5 is critical for PALB2-BRCA2 interaction [63].

Based on these premises, one non-coding event plus 28 frame-shift events predicted to disrupt expression of WD40 Δ-strand 7C were annotated as bona fide loss of function (LoF) events (Table 2). In-frame events (Table 1) were annotated as LoF if predicted to code for PALB2 protein isoforms lacking a functional domain, lacking a critical residue located in a functional domain, or targeting at least one WD40 Δ-strand. Otherwise, in-frame events were annotated as *uncertain coding potential*.

2.3.1 In-frame splicing events predicted to code for unstable or non-functional proteins (annotated as LoF):

-The splicing event Δ(E2) is predicted to code for a PALB2 isoform lacking 20 out of 35 coiled-coil residues, including residues that are critical for PALB2-BRCA1 interaction [65], [66]

-The splicing event Δ(E4) is predicted to code for a PALB2 isoform lacking the ChAM/DNA binding domain that is critical for PALB2 association with chromatin [67].

-The splicing event Δ6 is predicted to code for a PALB2 isoform lacking the WD40 β-strand 7D, that has a fundamental role in WD40 Δ-propeller toroidal folding [63]. Interestingly, clinical and functional data suggest that PALB2 Δ6 may code for an hypomorphic protein that retains some functionality, albeit with severely reduced stability [69].

-The splicing event Δ(E7) is predicted to code for a PALB isoform lacking the first blade. Only 7- or 8-blade WD40 Δ-propeller has been confirmed structurally, with 7-blade considered the most ideal Δ-sheet geometry [63], [68].

-The splicing event Δ(E9p30) targets β-strands 2C, and most of the linker region between blades 2 and 3 (linker region between β-strands 2C and 2D) [63], [70, p. 40].

-The splicing event Δ(E9) is predicted to code for a PALB2 isoform lacking the end of WD40 β-strand 2C, and β-strands 2D, 3A, 3B, and 3C [63], [70, p. 40]. This is probably incompatible with stable four-stranded antiparallel Δ-sheet blades 2 and 3.

-The splicing event Δ(E9_E10) is predicted to code for a PALB2 isoform lacking the end of WD40 β-strand 2C, the complete blade 3 (Δ-strands 3A, 3B, 3C, and 3D), and WD40 β-strands 4A, 4B, and 4C [63]. Similarly, the splicing event Δ(E10) is predicted to code for a PALB2 isoform lacking a stable blade 4. WD40 Δ-propeller blade 4 is directly involved in PALB2-BRCA2 interaction [63]. Further, both protein isoforms are predicted to lack residues A1025, critical for PALB2-BRCA2 interaction [63], and residue T1030, required for HR repair activity [68].

-The splicing event Δ(E11_E12) is predicted to code for a PALB2 isoform lacking Δ-strand 4D and blade 5 (Δ-strands 3A, 3B, 3C, and 3D). Similarly, the splicing event Δ(E12) is predicted to code for a PALB2 isoform lacking blades 5 (Δ-strands 5A, 5B, 5C, and 5D), 6 (Δ-strands 6A, 6B, 6C, and 6D), β-strands 7A, 7B, and 7C, and the critical residue Leu1143[63], [68]. WD40 Δ-propeller blades 4 and 5 are directly

involved in PALB2-BRCA2 interaction [63], and β-strand 7C is critical to seal the WD40 Δ-propeller toroidal folding [63].

2.3.2 Splicing events coding for PALB2 isoforms of uncertain stability and/or activity (annotated as uncertain):

-The splicing events ▼(E1q9) and Δ(E2p6) are predicted to introduce minor changes in the primary sequence of the coiled-coil domain (p.K16_L17ins3 and p.L17_K18del). The precise functional effect of inserting 3 residues in between K16 and L17, or deleting residues L17 and K18, if any, is unknown. Interestingly, the missense change K18R has no functional impact [66]

- The splicing event Δ(E5p24) is predicted to delete a protein region that does not include known functional domains and/or critical residues.

-The splicing events ▼(E7p42) is predicted to insert 14 residues (p.K862_N863ins14) in the linker region between Δ-strand 7D (residues 855-860) and Δ-strand 1A(residues 868-877) [63]. The splicing event Δ(10p3) is predicted to delete one of 3 consecutive glycines located in the linker region between blade 3 and blade 4(Δ-strands 3C and 3D) [63]. The possible impact of insertion/deletions at linker regions is difficult to predict, but cannot be assumed necessarily damaging. Nonetheless, WD40 Δ-propeller protein domains can tolerate insertions in between two consecutive blades, or in between two Δ-strands of the same repeat, as Bub3 structure shows [70].

2.4. Quantification of alternative splicing events by RNAseq: In brief, the ratio (reads supporting the alternative splicing event/reads supporting the corresponding reference transcript) is considered to provide quantitative information. For instance, Δ(E2p6) relative contribution to the expression is calculated as (total number of Δ(E2p6) supporting reads/total number of exon1-2 junction reads), and is expressed as a percentage. Similarly, Δ(E11) relative contribution to the expression is calculated as (total number of exon10-12 junction reads/(total number of exon10-11+exon11-12 junction reads/2)). The approach has been described previously [50]. Calculations were made independently for targeted RNAseq experiments (laboratory 1), whole-transcriptome RNAseq experiments (laboratory 2), and whole-transcriptome RNAseq experiments (laboratory 3). For each laboratory, we provide an average quantification pooling together data from all runs (all samples plus technical replicas).

*Quantification of alternative splicing events by digital PCR (dPCR)***:** The use of dPCR to quantify alternative splicing events has been described previously [71]. In brief, to quantify PALB2 Δ(E4_E5) transcripts, we used a FAM-labeled custom designed TaqMan assay (Applied Biosystems) specific for the E3-E6 junction (5'-AACACTCAG-ACTGAAACAGCAGAGC-3´). To quantify PALB2 E11-E12 junctions we used a FAM-labeled pre-designed TaqMan assay (Applied Biosystems, Hs00226617_m1) specific for that junction (5'-TTCTGAAATG-GGGCTTCTCTTTATT-3'). As a reference we used a 2'-chloro-7'phenyl-1,4-dichloro-6-carboxy-fluorescein labeled (VIC-labeled) pre-designed TaqMan assay (Applied Biosystems,

Hs00954119_m1) specific for the PALB2 E4-E5 junction (5'-GTGAAAG-GGAAGAAAAGTCGTCATC-3'). Relative quantification experiments were performed combining FAM and VIC assays in individual 20k chips. All experiments were performed on a QuantStudio 3D Digital PCR 20K platform according to the manufacturer's instructions (Applied Biosystem, Foster City, CA).

Data was analyzed in the QuantStudio 3D Analysis Suit Cloud Software v2.0 (Applied Biosystem, Foster City, CA), defining FAM as Target. Default settings were used in all cases. After reviewing automatic assessment of the chip quality by the proprietary software, only green (data meets all quality thresholds, review of the analysis result not required) and yellow flag chips (data meets all quality thresholds, but manual inspection is recommended) were considered for further analyses. We used FAM/VIC ratios as proxies for E4E5 exclusion rate and E11E12 inclusion rate.

3. PVS1 status (warranted vs. not warranted) for every possible PTC-NMD and splice site variant at the PALB2 locus.

Only high-confidence alternative splicing events were considered for this analysis.

3.1 PVS1 status of PALB2 PTC-NMD variants: For this analysis we consider the possibility that naturally occurring alternate gene transcripts may result in skipping of truncating variants within specific exons and resulting in reduced or no haplo-insufficiency (*rescue transcripts*). Only *PALB2* alternate gene transcripts not annotated as LoF (see coding potential) are plausible rescue transcripts. Further, a plausible *rescue transcripts* must reach a certain expression level threshold (for the purpose of this study, the threshold is arbitrarily set at a very conservative 10% of the corresponding reference transcript according to targeted RNAseq experiments). If no plausible rescue transcript is predicted for a particular PTC-NMD variant, PVS1 is warranted. If at least one candidate rescue transcript is predicted, PVS1 is not warranted.

3.2 PVS1 status of PALB2 splice site variants: To decide if PVS1 is warranted for every possible splice site variant at the *PALB2* gene, we made predictions on possible outcomes for these alterations. Depending on the availability of RNA splicing assays in the scientific literature, splice site predictions were based on the following assumptions:

(i) We have identified splicing assays performed in carriers of five *PALB2* splice site variants (c.48G>A, c.2515-1G>T, c.2835-1G>C, c.3113G>A, and c.3113+5G>A), that we complemented with splicing assays performed in *PALB2* c.2748+2T>G and c.3350+1G>A carriers (Supplemental Table 4 and Supplemental Figure 6). We used data on these seven assays to predict the most likely outcome of splice site variants located at 3 *PALB2* donor sites (exons 1, 7 and 12), and two *PALB2* acceptor sites (exons 6 and 9). For instance, alternate gene transcripts predict up to 3 possible RNA products for splice site variants impairing the exon 1 donor site: Δ(E1q17), Δ(E1q169) and/or ▼(E1q9). However, an existing RNA splicing assay shows that c.48G>A, a genetic variant impairing exon 1 donor site, causes only Δ(E1q17). Based on this study, we predict that any

splice site variant targeting exon 1 donor site will produce Δ(E1q17), a LoF transcript. For that reason, we consider that PVS1 is warranted for any splice site variant at exon 1 donor site.

(ii) If we have not identified RNA splicing assays in the scientific literature, we predicted that splice site variants may produce both exon skipping and/or up-regulation of naturally occurring alternate gene transcripts (not hampered by the splice site variant under consideration). According to the coding potential, these predicted products are annotated as bona fide LoF or uncertain coding potential. For the purpose of this study, PVS1 is warranted only if all predicted products are bona fide LoF transcripts. Otherwise, PVS1 is not warranted. For instance, we have not identified in the scientific literature RNA splicing assays performed in carriers of splice site variants targeting the exon 2 acceptor site. Therefore, we predict that variants targeting this splice site may produce Δ(E2)+▼(E2p26) and/or Δ(E2p6) transcripts. Both Δ(E2) and ▼(E2p26) have been annotated as bona fide LoF transcripts, but Δ(E2p6) has been annotated as a transcript of *uncertain* coding potential. Due to the latter, we consider that PVS1 is not warranted for splice site variants targeting exon 2 acceptor site (unless RNA splicing assays prove otherwise). We have used this approach to predict the most likely outcome of splice site variants located at 9 *PALB2* donor sites (exons 2, 3, 4, 5, 6, 8, 9, 10, and 11), and 10 *PALB2* acceptor sites (exons 2, 3, 4, 5, 7, 8, 10, 11, 12, and 13).

4. Splicing analysis of *PALB2* variants c.212-1G>A (IVS3-1G>A) and c.1684+1G>A (IVS4+1G>A) using a reporter minigene.

4.1 Minigene Construction: Minigene mgpb2_ex4-6 was assembled in two steps. First, exons 5 and 6 were amplified with Phusion High Fidelity Polymerase (Thermo Fisher Scientific, Waltham, MA, USA) and primers 5' GGTGGCGGCCGCTCTAGAACTAGTGGATCCCCCGGAGTCATGGATGGGAAAAGTAA 3‘ and 5' GACGGTATCGATAAGCTTGATATCGAATTCCTGCATTGGCATAGAAACTTTAAGAGG 3‘ (1,741 bp) under standard conditions, and then this fragment was inserted into the pSAD vector [72], [73] by Overlapping Extension PCR [74]. Second, exon 4 and its flanking intronic sequences were amplified with Phusion High Fidelity Polymerase and primers 5' TATATATCTAGAGTTAAGAGAAGAGATTGTGTGA 3' and 5' TATATAGGATCCATACATTTTCCTTTTCAGTGTT 3' (2,623 bp). This insert was cloned upstream exon 5 of the previous construct between restriction sites XbaI and BamHI. Structure of the final construct (mgpb2_ex4-6): XbaI- Ivs3 (416 bp) – ex4 (1473 bp) – ivs4-1 (734 bp) – BamHI – [overlapping] - ivs4-2 (266 bp) – ex5 (830 bp) – ivs5 (364 bp) – ex6 (72 bp) – ivs6-1 (209 bp) - [overlapping].

DNA variants c.212-1G>A and c.1684+1G>A were introduced by site-directed mutagenesis with the QuikChange Lightning kit (Agilent, Santa Clara, CA). Mutant clones were confirmed by sequencing (Macrogen, Madrid, Spain)

4.2 Transfection and RT-PCR: Approximately 4x105 MCF7 cells were grown to 90% confluence in 0.5 mL of medium (MEME, 10% Fetal Bovine Serum, 2 mM glutamine, 1% Non-essential amino acids and 1% Penicillin/Streptomycin) in 4-well plates (Nunc, Roskilde, Denmark). Cells were transiently transfected with 1 μg of each minigene and 2 μL of low toxicity Lipofectamine (Life Technologies, Carlsbad, CA). To inhibit nonsense mediated decay (NMD), cells were incubated with cycloheximide (Sigma-Aldrich, St. Louis, MO) 300 μg/mL for 4-5 hours. RNA was purified with the Genematrix Universal RNA Purification Kit (EURx, Gdansk, Poland) with on-column DNAse I digestion to degrade genomic DNA that could interfere in RT-PCR.

Retrotranscription was carried out with 400 ng of RNA and RevertAid H Minus First Strand cDNA Synthesis Kit (Life Technologies), using gene specific primer RTPSPL3-RV (5'TGAGGAGTGAATTGGTCGAA 3'). Samples were incubated at 42ºC for 1 hour, and reactions were inactivated at 70ºC for 5 min. Then, 1-2 μl of the resultant cDNA were amplified with SD6-PSPL3_RTFW (5'-TCACCTGGACAACCTCAAAG-3') and RTpSAD-RV (Patent P201231427, CSIC) (expected canonical transcript: 2,556 nt), using Platinum Taq DNA polymerase (Life Technologies). Samples were denatured at 94ºC for 2 min, followed by 35 cycles consisting of 94ºC for 30 sec, 59ºC for 30 sec, and 72ºC (1 min/kb), and a final extension step at 72ºC for 5 min. RT-PCR products were sequenced by Macrogen (Madrid, Spain).

5. RT-PCR analysis in carriers of *PALB2* variants c.2748+2T>G and c.3350+1G>A

RT-PCR analysis were performed from 200 ng of total RNA using the Onestep RT-PCR kit (Qiagen) as previously described [75] on patients Paxgene blood RNA samples, obtained from patients with HBOC syndrome. Reactions were performed in two different RT-PCR reaction, with specific primers (available on request) encompassing the variant as previously preconized [5] RT-PCR products were separated on a 1.5% agarose gel.

6. Whole-gene CloneSeq analysis in carriers of *PALB2* variants c.3133+5G>C (IVS10+5G>A), c.3350+4A>C (IVS12+4A>C), and c.3350+5G>A (IVS12+5G>A)

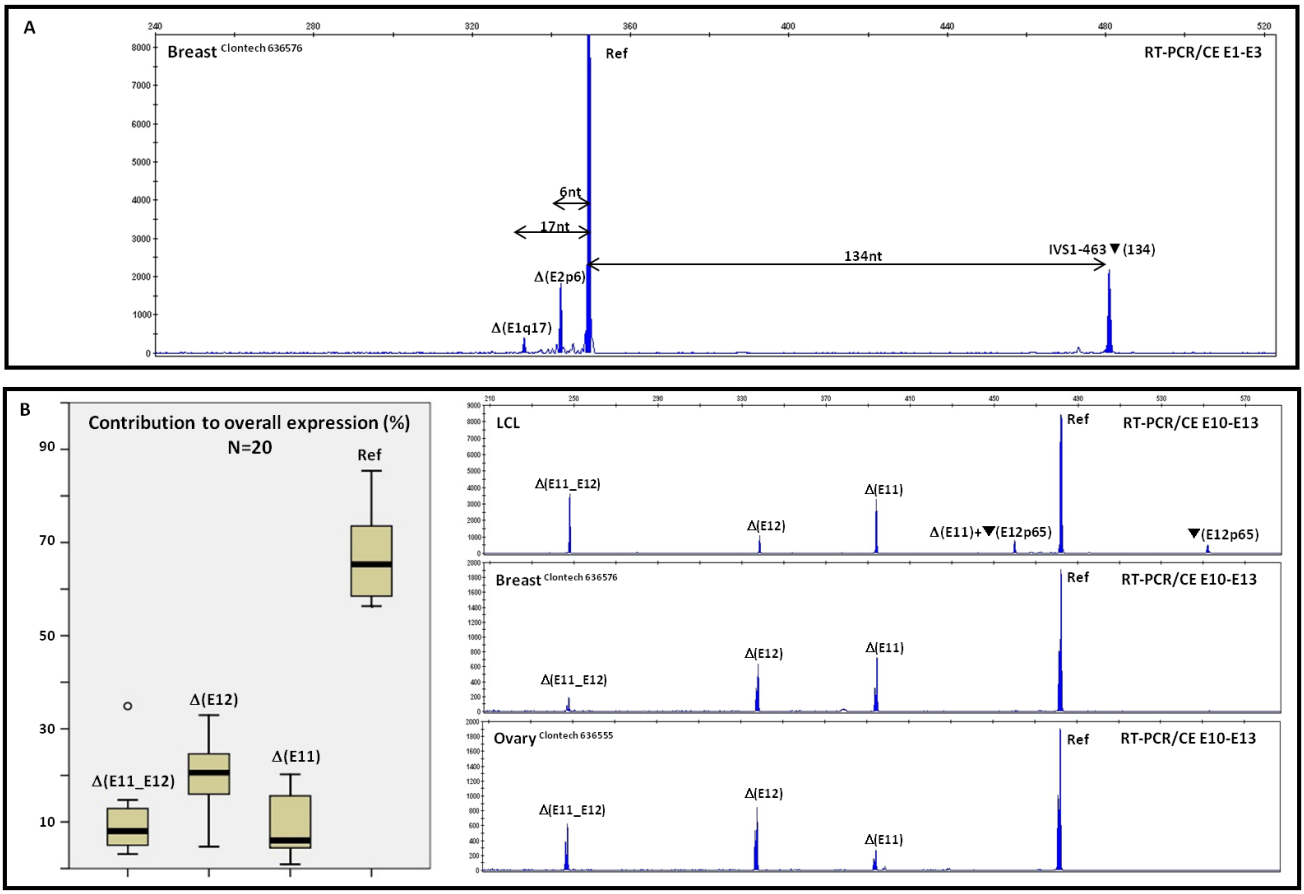
Blood from carriers participating in the Ambry Genetics Family Studies program was drawn in PAXgene Blood RNA Tubes and stored according to the manufacturer's recommendations (PreAnalytiX, Hombrechtikon, Switzerland). RNA was extracted using the PAXgene Blood RNA Kit according to the recommended protocol (PreAnalytiX)

2. Supplemental Tables

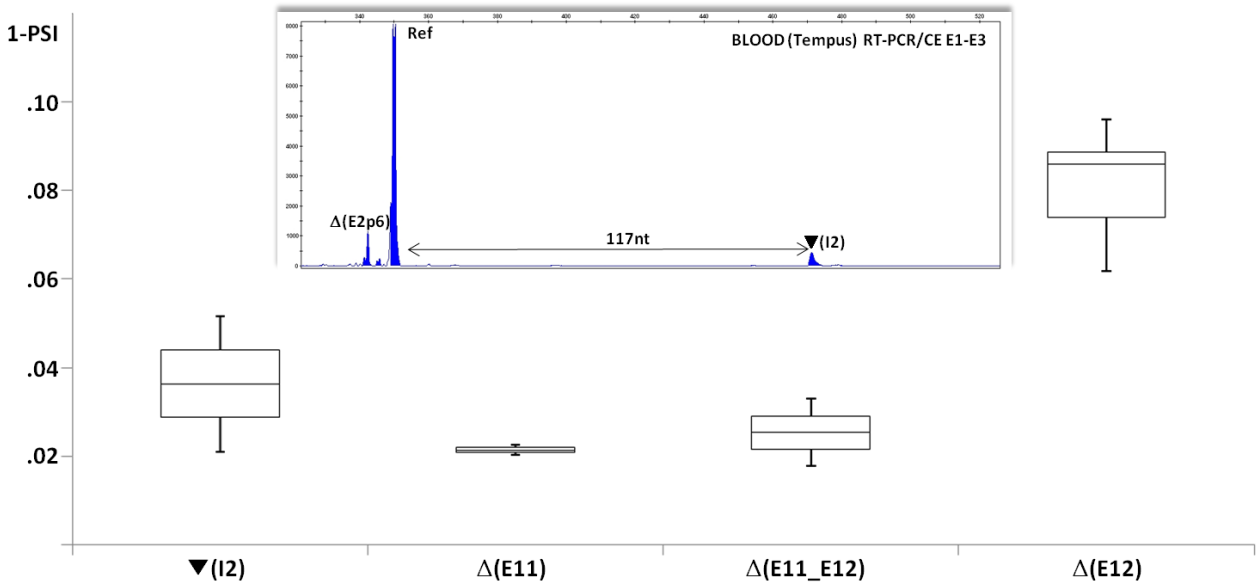
PALB2 Splice Site	Possible mRNA Products (up-regulation of alternate gene transcripts)		RNA Splicing Assays		
	Coding potential LoF	Coding potential uncertain	Tested Variant	Observed Product	ref
E1 donor	Δ(E1q17),Δ(E1q169)	▼(E1q9)	c.48G>A	Δ(E1q17)	[76]
E2 acceptor	Δ(E2)	Δ(E2p6)	c.49-1G>A	Δ(E2p6)	Unpublished data
E4 acceptor	Δ(E4),Δ(E4_E5)	-	c.212-1G>A	Δ(E4_E5)	Supplemental Figure 6A
E4 donor		-	c.1684+1G>A		Supplemental Figure 6A
E6 acceptor	Δ(E6)	-	c.2515-1G>T	Δ(E6)	[77]
E6 donor		-	c.2586+1G>A		[69]
E9 acceptor	Δ(E9),Δ(E9_E10)	Δ(E9p30)	c.2835-1G>C	Δ(E9p30) and Δ(E9)	[77]
E10 donor	Δ(E9_E10),(E10),Δ(E10q31)	-	c.3113G>A	Δ(E10q31) and Δ(E10)	[77]
			c.3113+5G>A	Δ(E9_E10)	[64]
			c.3113+5G>A	Δ(E10q31)	Supplemental Figure 6C
E12 donor	Δ(E11_E12),Δ(E12)	-	c.3350+1G>A	Δ(E12)	Supplemental Figure 6B
			c.3350+4A>C	Δ(E11_E12), Δ(E12)	Supplemental Figure 6C
			c.3350+5G>A		Supplemental Figure 6C

Supplemental Table 4. We have identified in the scientific literature five splicing assays performed in carriers of *PALB2* variants impairing two donor sites (exons 1 and 10) and two acceptor sites (exons 6 and 9). On preparing this manuscript, we learnt of a Greek family carrying a *PALB2* germ-line variant (c.49-1G>A) impairing *PALB2* exon 2 acceptor site. A splicing assay revealed that the variant does not cause exon 2 skipping, but up-regulation of Δ(E2p6), thus supporting our predictions (Dr. Georgios Tsaousis, Genekor Medical S.A., personal communication). Further, we have performed splicing assays in carriers of *PALB2* variants impairing one acceptor site (exon 4) and four donor sites (exons 4, 7 10, and 12). Overall, all these studies support the hypothesis that naturally occurring alternate gene transcripts provide predictive information on the possible outcome of splice site variants, in particular predictions on the possible up-regulation of cryptic sites usage (Δ(E1q17),□Δ(E2p6), Δ(E9p30), Δ(E10q31)), and/or multi-cassette events (Δ(E4_E5), Δ(E9_E10), Δ(E10_E11)) . Similar observations have been made previously for *BRCA1* and *BRCA2* splice site variants. For instance, variants impairing *BRCA1* exons 5 and 9 donor sites, *BRCA2* exons 6 and 13 donor sites, and *BRCA2* exon 23 acceptor site produce respectively Δ(E5q22), Δ(E9_E10), Δ(E5_E6), Δ(E12_E13), and Δ(E23p51) transcripts, all them described as alternate gene transcripts in control samples [52], [78], [79].

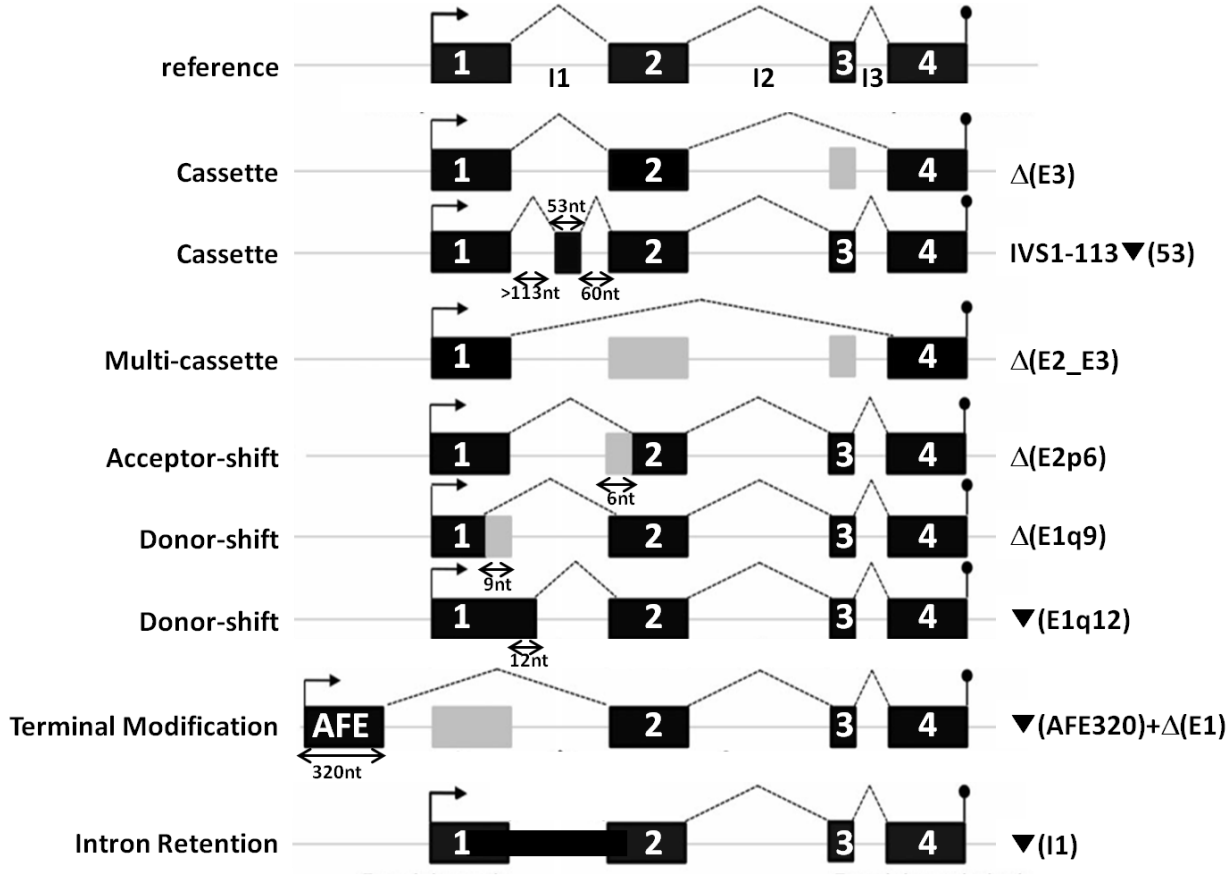
3. Supplemental Figures



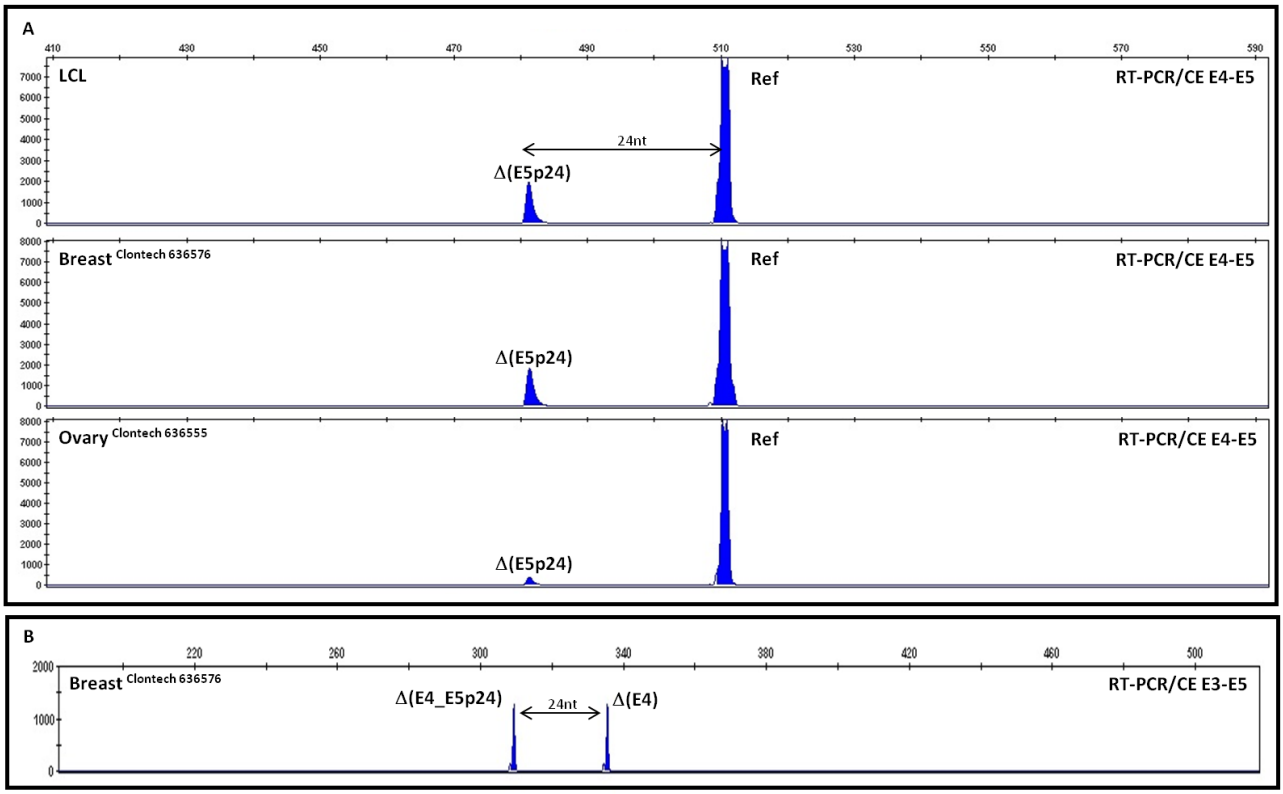
Supplemental Figure 1A. ***PALB2*** **alternative splicing analysis by RT-PCR/CE. Panel A** show a representative example of one RT-PCR/CE experiment performed with a forward primer located in exon 1 and a Fam-labeled reverse primer located in exon 3. The experiment was performed using commercially available RNA from non-malignant healthy breast tissue. Once the PCR product (peak) corresponding to the reference transcript (ref) is identified, the identity of the remaining peaks is imputed from size differences, as indicated. Splicing events Δ(E1q17), Δ(E2p6), and IVS1-463▼(134) are supported as well by RNAseq experiments. **Panel B**, left. The boxplot (minimum, Q1, median, Q3 and maximum values are displayed) show the average contribution of Δ(E11), Δ(E12), Δ(E11_E12), and reference transcripts to the overall expression as determined by exon10-exon 13 RT-PCR/CE experiments (13 independent LCLs, 5 technical replicas of Breast Clontech, and two technical replicas of Ovary Clontech). See methods for further details. Representative examples of individual RT-PCR/CE experiments performed in LCLs, Breast, and Ovary are shown to the right. Due to their very low contribution to the overall expression, splicing events Δ(E11)+▼(E12p65) and ▼(E12p65) are not detected in all technical replicas (stochastic amplification of low expressed transcripts). Yet, both events have been detected by RT-PCR/CE and RNAseq experiments.



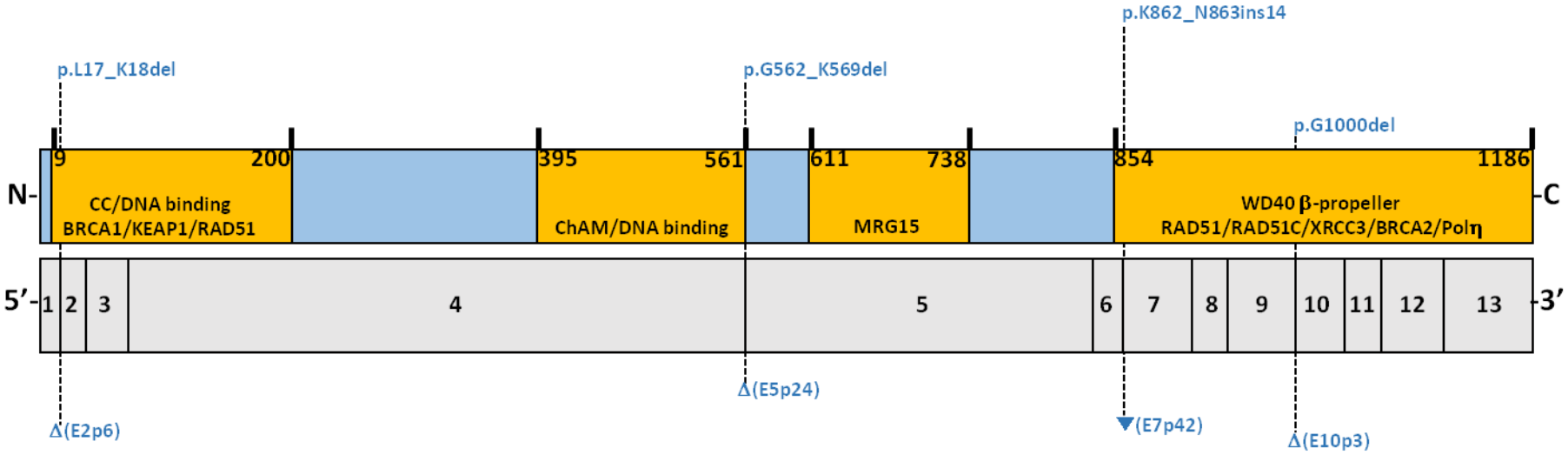
Supplemental Figure 1B. *PALB2* whole-gene CloneSeq alternative splicing analysis. The analysis performed in 5 control samples (PAXgene RNA) identified only 4 *PALB2* alternative splicing events representing on average ≥2.5% of the corresponding reference transcript (1-PSI≥0.025). whole-gene CloneSeq alternative splicing analysis coincides with targeted RNAseq, whole-transcriptome RNAseq, and RT-PCR/CE in identifying splicing events Δ(E11), Δ(E12), and Δ(E11_E12) among the top expressed *PALB2* splicing events. Interestingly, CloneSeq adds to the top list the full retention of *PALB2* intron 2 (with 117nt, this is the shortest reference intron). This event is not detectable by the RNAseq analysis pipelines used in this study but has been detected by RT-PCR/CE, albeit the data suggest a very low contribution to the overall expression.



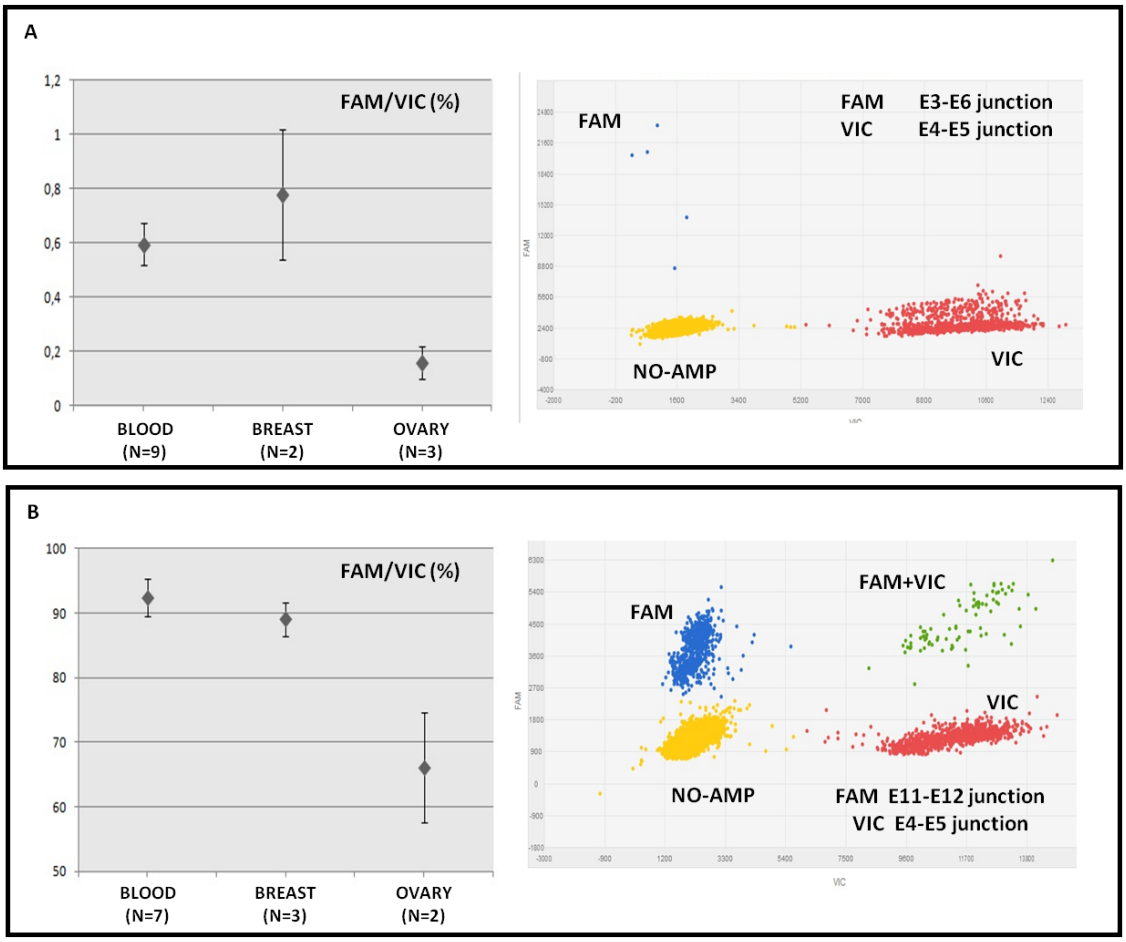
Supplemental Figure 2. We designated alternative splicing events with a code that combines the following symbols: E (reference exon), I (reference intron), Δ(skipping of reference exonic sequences), ▼(inclusion of reference intronic sequences), p (acceptor shift), q (donor shift), and IVS (novel exon in between two reference exons). The figure shows a theoretical four exons gene (reference) in which eight alternative splicing events have been described: i) Δ(E3), or skipping of reference exon 3, ii) IVS1-113▼(53), or inclusion of a 53 nucleotide alternative exon in between reference exons 1 and 2 iii) Δ(E2_E3), or skipping of consecutive exons 2 and 3, iv) Δ(E2p6), or a 6 nucleotide skipping (acceptor shift) in exon 2, v) Δ(E1q9), or a 9 nucleotide skipping (donor shift) in exon 1, vi) ▼(E1q12), or a 12 nucleotides inclusion (acceptor shift) in exons 1,vii) ▼(AFE320)+Δ(E1), or use of an alternative first exon of 320 nnucleotides, and viii) ▼(I1), or retention of intron 1. If the splicing event involves both skipping of reference exonic sequence and inclusion of reference intronic sequence we have used the general scheme Δ+▼. For instance, an alternative splicing event combining a donor shift skipping 31 nucleotides of exon 10 with an acceptor shift adding 23 nucleotides to exon 11 is described as Δ(E10q31)+▼(E11p23).



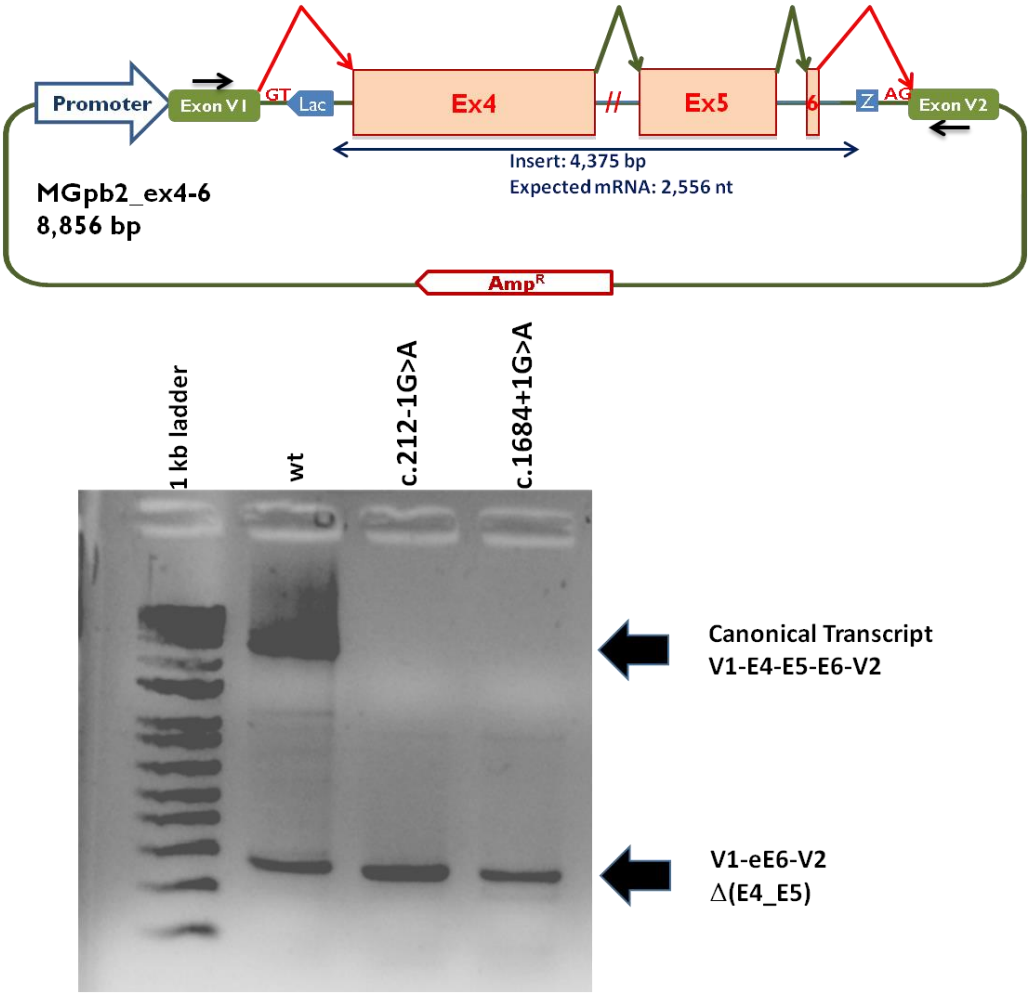
Supplemental Figure 3. Despite the fact that is not supported by RNAseq data, we have annotated the exon 5 acceptor shift Δ(E5p24) as a high-confident event. RT-PCR/CE data seems consistent, with a compatible peak detected in BLOOD, BREAST, and OVARY (panel A) with various primer combinations (Supplemental Table 3). Further on, the event has been inferred also in combination with Δ(E4) (Panel B and Supplemental Table 2).



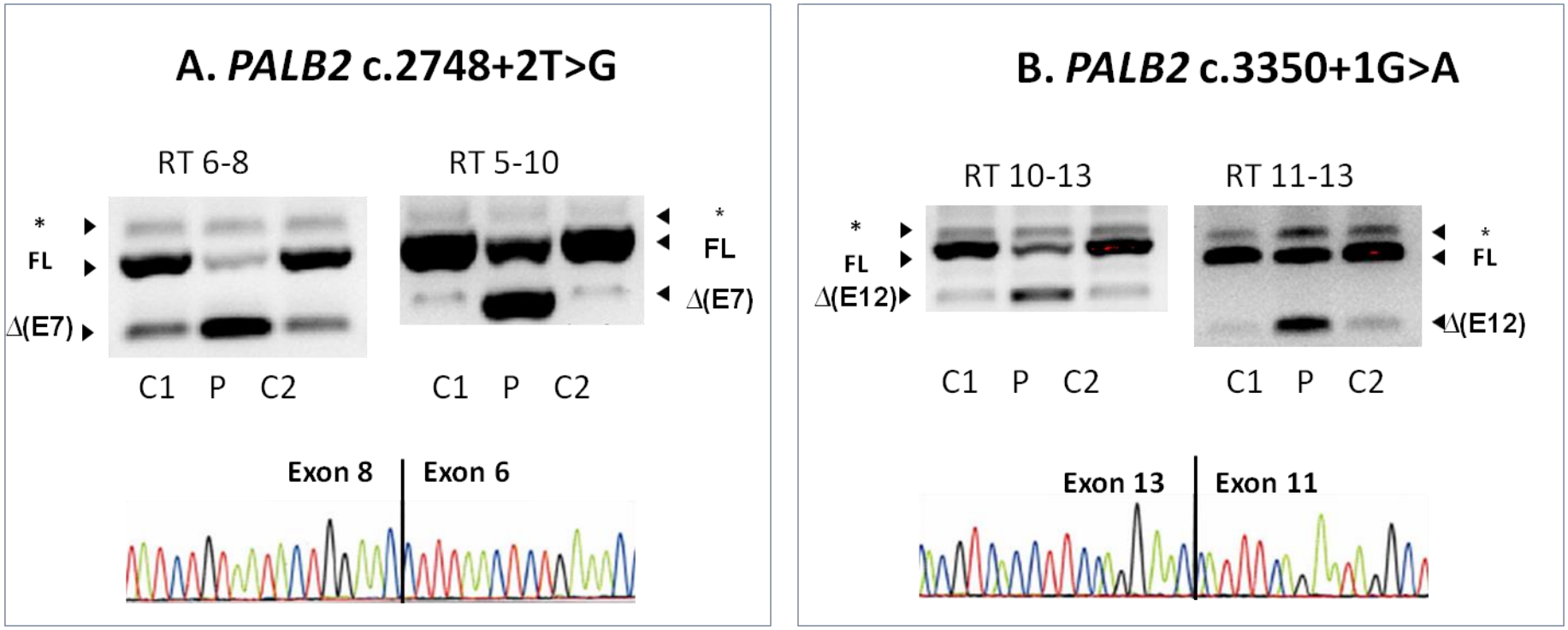
Supplemental Figure 4. *PALB2* functional domains and alternative splicing events. Schematic representation of the PALB2 protein (top) and reference mRNA (bottom) that shows the correspondence between functional protein domains (orange) and coding exons (grey). Residues 9 to 200 code for one DNA binding domain (spanning the whole region), Coiled-Coil (CC) and BRCA1 interacting (residues 9-44), KEAP1 binding (88-94), and RAD51 binding (101-184) domains. Residues 395 to 561 code for a chromatin associated motif (ChAM, residues 395-446), and a second DNA binding domain (residues 446-561). Residues 611 to 738 code for one MRG15 binding domain. Finally, C-terminal residues 854 to 1186 code for the RAD51-, RAD51C-, XRCC3-,BRCA2-, and Polη-interacting WD40 Δ-propeller domain [60]–[62]. The PALB2 WD40 Δ-propeller domain is composed of 7 WD40 repeats. Based on this information, we annotated alternative splicing events as LoF (predicted to code for a non-functional or unstable protein) or *uncertain* (the possibility of a functional or partially functional protein product must be considered). Due to its potential relevance for the classification of splice site and PTC-NMD variants, the figure shows the predicted coding effect (top) of 5 naturally occurring alternative splicing events (bottom) that we annotate as *uncertain*.



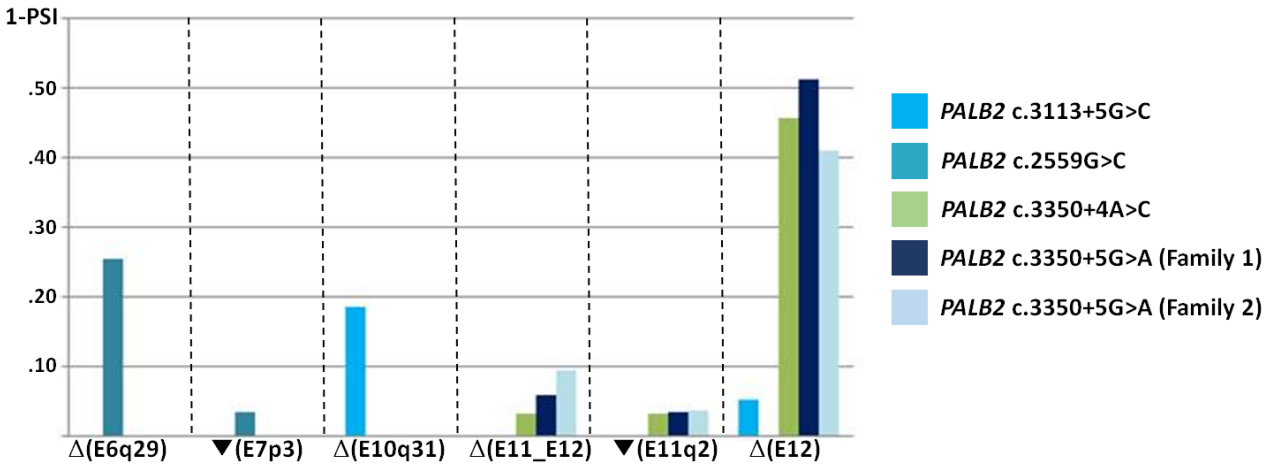
Supplemental Figure 5. Digital PCR quantification of *PALB2* alternative splicing events. Panel A left shows the *PALB2* Δ(E4_E5)(FAM)/E4-E5(VIC) expression ratio as determined by digital PCR in BLOOD (N=9 independent samples, 7 LCLs plus 3 Tempus samples), BREAST (N=2 technical replicas of Breast Clontech), and OVARY (N=3 technical replicas). The data shows that the contribution of Δ(E4_E5) to the *PALB2* expression level is rather low (<1%) if compared with the corresponding reference transcript containing exon4-5 junctions. **Panel B** left shows the *PALB2* E11-E12 (FAM)/E4-E5(VIC) expression ratio as determined by digital PCR in BLOOD (N=7 independent samples, 3 LCLs plus 4 Tempus samples), BREAST (N=3 technical replicas of Breast Clontech) and OVARY (N=2 technical replicas Ovary Clontech). The data indicates that the expression level of reference exon 11-12 junctions is lower than the corresponding level of exon 4-5 junctions (8%-35% lower, depending on the sample analyzed) supporting the finding that alternative splicing events skipping exon 11, exon 12, or both, make a significant contribution to the *PALB2* expression in BLOOD, BREAST, and OVARY. Panel B right shows a representative example of a digital PCR experiment with combining in a single chip a FAM-labeled TaqMan assay recognizing *PALB2* exon 11-12 junctions, and a VIC-labeled assay recognizing exon 4-5 junctions.



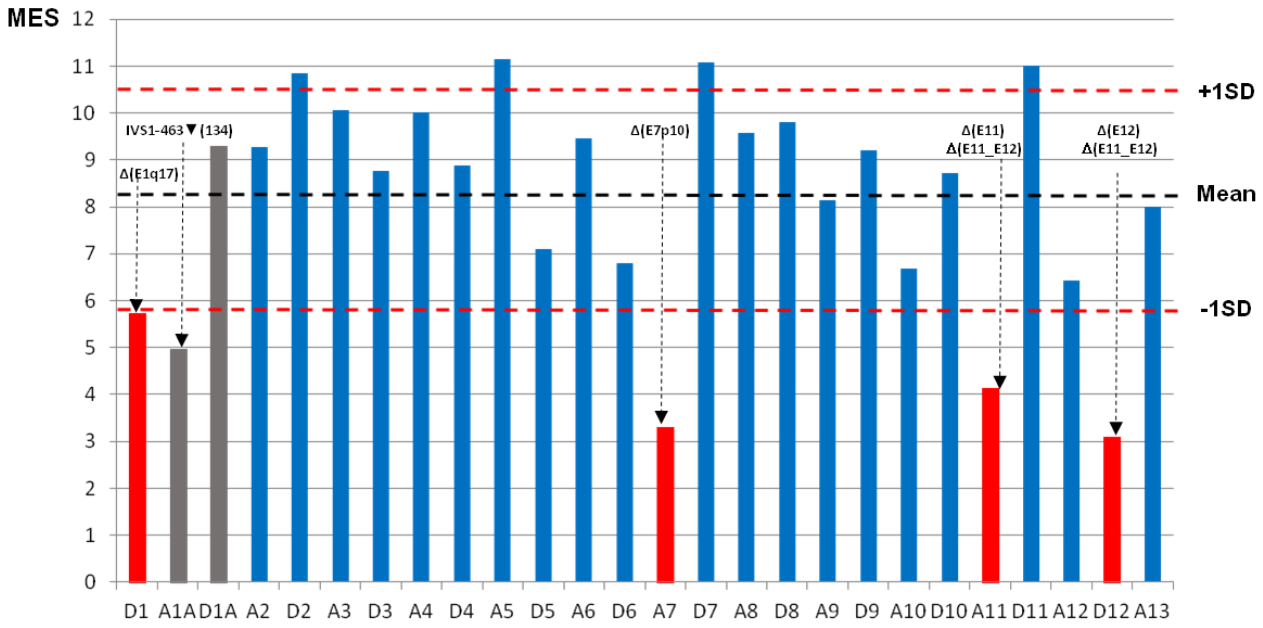
Supplemental Figure 6A. Reporter minigene MGpb2_ex4-6 indicates that the splice site variants *PALB2* c.212-1G>A (IVS3-1G>A) and c.1684+1G>A (IVS4+1G>A) do not cause exon 4 skipping, but skipping of exons 4 and 5 together. Remarkably, *PALB2* Δ(E4_E5) is a high-confidence alternative splicing event detected both in control samples and the wild-type minigene.



Supplemental Figure 6B. RNA splicing assays performed in carriers of *PALB2* splice site variants c.2748+2T>G and c.3350+1G>A. P: Patient, C: controls, *heteroduplex, FL: full length transcript, Δ: skipping of exon A. *PALB2* variant c.2748+2T>G, RT-PCR were done with forward primer in exon 5 and the reverse primer in exon 10 for one and with forward primer in exon 6 and the reverse primer in exon 8 for the second. The reverse sequence refers to the product obtained from the lower band. B. *PALB2* variant c.3350+1G>A, RT-PCR were done with forward primer in exon 10 and the reverse primer in exon 13 for one and with forward primer in exon 11 and the reverse primer in exon 13 for the second. The reverse sequence refers to the product obtained from the lower band



Supplemental Figure 6C. CloneSeq splicing analysis performed in carriers of four *PALB2* genetic variants. *PALB2* c.2559G>C (located in exon 6), creates a *de novo* donor site causing skipping of 29 nucleotides. *PALB2* c.3113+5G>C (IVS10+5G>A) impairs exon 10 donor site. Interestingly, the detected outcome is not exon 10 skipping, but up-regulation of a cryptic donor site already detected in control samples. *PALB2* c.3350+4A>C (IVS12+4A>C) and c.3350+5G>A (IVS12+5G>A) impairs exon 12 donor site, causing exon 12 skipping, but also skipping of exons 11 and 12 together. Interestingly, Δ(E11_E12) has been detected in control samples.



Supplemental Figure 7. *PALB2* canonical splice sites and alternative splicing. The chart shows in red and blue the MaxEnt Score (MES) for all donor (D) and acceptor (A) sites involved in the production of the reference transcript ENST00000261584 . Dashed lines indicate mean and ±1 Standard Deviation of the 24 MES Scores at *PALB2* canonical splice sites. Scores <-1SD are highlighted in red. MaxEnt Scores for the donor and acceptor sites of alternative exon IVS1-463▼(134) are shown in grey. All 6 splicing events representing >1% of the reads supporting the corresponding reference transcript are represented as well in the chart. Apparently, Δ(E1q17) is explained by a weak (<-1SD) donor site in exon 1, Δ(E7p10) by a weak acceptor site in exon 7, Δ(E11) by a weak acceptor site in exon 11, and Δ(E12) by a weak donor site in exon 12. Δ(E11_E12) can be explained by both a weak acceptor site in exon 11 and a weak donor site in exon 12. Interestingly, the chart suggest that the low inclusion rate of IVS1-463▼(134) is due to a weak acceptor site. Yet, it is plausible that rare genetic variants located close to *PALB2* c.49-463 improve the canonical acceptor site, causing IVS1-463▼(134) inclusion (a PTC-NMD alteration) in most transcripts, thus providing a putative disease mechanism for deep intronic variants in *PALB2* intron 1. A similar mechanism has been described for the deep intronic variant *BRCA2* c.6937+594T>G in French families [80], albeit recent studies have challenge the initial pathogenic classification of this BRCA2 variant [81].

VI. ANNEXE F : protocole utilisé pour la capture RNA-seq *long read*

Trousses de réactifs utilisées

Trousses	Vendeur (ref)
SMARTer PCR cDNA Synthesis kit	Clontech (634925)
PrimeSTAR GXL DNA polymerase	Clontech (R050A)
Tris EDTA buffer solution (100X)	SIGMA (T9285)
Takara LA Taq DNA Polymerase Hot Start version	Clontech (RR042A)
SMRTbell library construction and sequencing	PacBio
Blocker polyT Oligo (5‘ TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT / 3invdT/3’)	Eurogentec (1mM)
Blocker SMARTer PCR Oligo (5‘ AAG CAG TGG TAT CAA CGC AGA GTA 3’)	Eurogentec (1mM)
AMPure XP for PCR Purification	Beckman (A63881)
SureSelectXT Target Enrichment box 1/box 2	Agilent (5190-4394/5190-6261)
Dynabeads MyOne Streptavidin T1	Invitrogen (65602)

Échantillons utilisés

N° échantillon	Description	[ARN, ng/µL] (RIN)
Ech 1	Lignée lymphobalstoïde	420 (8.8)
Ech 2	Leucocytes stimulés – puromycine	718 (9.9)
Ech 3	Leucocytes stimulés + puromycine	550 (9.5)
Ech 4	Lignée c.4063+3A>T	500 (8.8)

1. Reverse transcription

La quantité d’ARN total utilisée a été fixée à 1 µg.

Préparation du mélange réactionnel pour la ligation des ARN avec le CDS Primer IIA, dans une barrette PCR.

Réactifs	Volume (pour 1 ech)
ARN total	1 – 3 µL
3’ SMART® CDS Primer IIA	1 µL
QSP eau sans nucléase	4.5 µL

Après mélange et rapide centrifugation, placer la barrette PCR dans le thermocycleur et lancer le programme suivant : 72°C pendant 3 min puis diminution en escalier par pas de 0.1°C/sec jusqu’à 42°C. Puis laisser les échantillons à 42°C.

Pendant cette étape, préparer le mélange pour la reverse transcription :

Réactifs	Volume (pour 1 ech)
5X First-Strand Buffer	2 µL
DTT (100 mM)	0.25 µL
dNTP (10 mM)	1 µL
SMARTer IIA Oligonucleotide (12 µM)	1 µL
RNase Inhibitor	0.25 µL
SAMRTScribe Reverse Transcriptase*	1 µL
Volume total	**5.5 µL**

* Ajouter l'enzyme au dernier moment

Préparer le mélange en fonction du nombre d'échantillon, chauffer le 1 min à 42°C. Distribuer 5.5 µL du mélange à chaque échantillon, en limitant l'évaporation. Mélanger par aspiration-refoulement et centrifuger les rapidement. Lancer le programme : 42°C pendant 90 min et finir par 70°C pendant 10 min.

Diluer les 10 µL d'ADNc (4.5+5.5) dans 190 µL de tampon Tris EDTA (1X). A noter que le tampon Tris EDTA est fourni concentré à 100X, donc à diluer dans de l'eau sans nucléase pour avoir 1X.

Les échantillons peuvent être stockés à -20°C.

2. PCR optimisation

Il est hautement recommandé de procéder à une optimisation du nombre de cycle de PCR pour réduire au maximum le risque de biais de PCR.

Préparation du mélange de PCR :

Réactifs	Volume (pour 1 ech)
5X PrimeSAR GXL buffer	10 µL
ADNc dilué	10 µL
dNTP mix (2.5 mM chacun)	4 µL
5' PCR Primer IIA (12 µM)	1 µL
Eau sans nucléase	24 µL
PrimeSTAR GXL DNA Polymerase	1 µL
Volume total	**50 µL**

Pour tester un intervalle de 10 à 14 cycles, utiliser les programmes :

Dénaturation initiale :

- 98°C pour 30 sec

10 cycles :

- 98°C pour 10 sec
- 65°C pour 15 sec
- 68°C pour 10 min

Extension finale :

- 68°C pour 5 min

A la fin du programme prélever 5 µL du mélange réactionnel. Sur les 45 µL restant lancer le programme :

2 cycles :

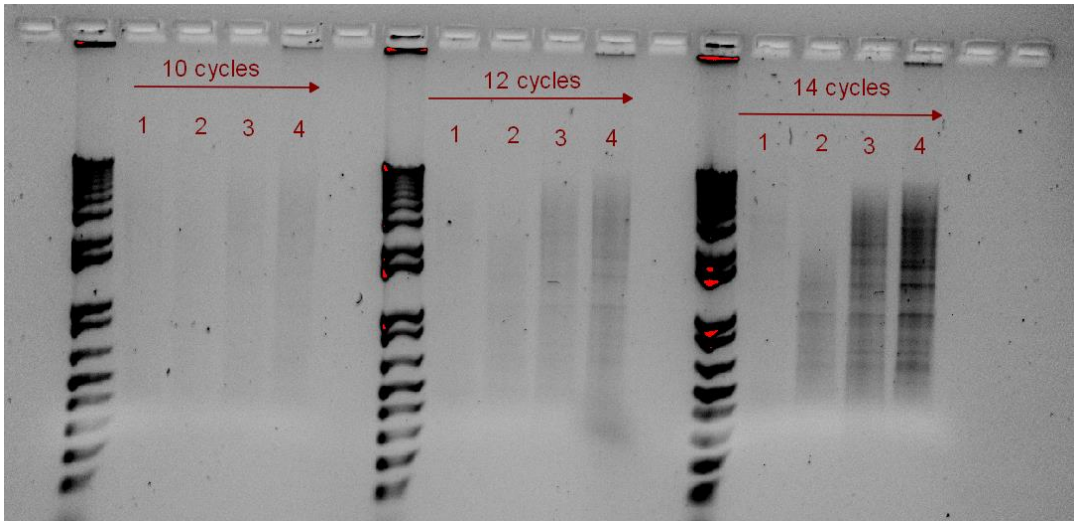
- 98°C pour 10 sec
- 65°C pour 15 sec
- 68°C pour 10 min

Extension finale :

- 68°C pour 5 min

A la fin du programme prélever 5 µL du mélange réactionnel. Sur les 40µL restant répéter le même programme pour un total de 14 cycles.

Puis faire migrer les différents prélèvements sur un gel d'agarose à 1 %. Exemple du gel obtenu lors de notre expérimentation :



Ici, 12 cycles se montrent optimaux pour amplifier les ADNc.

3. PCR à large échelle

Une fois le nombre optimal de cycles déterminé, préparer la PCR à large échelle avec 16 réactions par échantillon. Ici pour 4 échantillons, 64 (4 x 16) puits de PCR ont été utilisés.

Préparer le mélange réactionnel :

Réactifs	Volume (1 réaction)	Volume (1 ech)	Volume (4 ech)
5X PrimeSAR GXL buffer	10 µL	160 µL	660 µL
ADNc dilué	10 µL	160 µL	16 x 10 µL
dNTP mix (2.5 mM chacun)	4 µL	64 µL	264 µL
5' PCR Primer IIA (12 µM)	1 µL	16 µL	66 µL
Eau sans nucléase	24 µL	384 µL	1584 µL
PrimeSTAR GXL DNA Polymerase	1 µL	16 µL	66 µL
Volume total	**50 µL**	**800 µL**	**2640 µL**

Distribuer 64 x 50 µL dans des puits de PCR puis lancer le programme :

Dénaturation initiale :

- 98°C pour 30 sec

X cycles :

- 98°C pour 10 sec
- 65°C pour 15 sec
- 68°C pour 10 min

Extension finale :

- 68°C pour 5 min

4. Purification des produits de PCR à large échelle

Avant de procéder à la purification des produite de PCR, il est impératif d'éliminer un contaminant des billes AMPure XP qui interagit avec la liaison ADN et polymérase de la réaction de séquençage.

Pour 1 mL de billes AMPure XP :

1. Placer les billes sur un séparateur magnétique
2. Récupérer le PEG surnageant (ne surtout pas jeter)
3. Sur les billes ajouter 1 mL d'eau sans nucléase
4. Remettre les billes en suspension en vortexant de 30 à 60 sec
5. Placer les billes sur le séparateur magnétique et jeter le surnageant
6. Répéter les étapes 3 à 5 pour un total de 5 lavages
7. Sur les billes ajouter 1 mL de tampon d'élution, ici tampon EB de Qiagen (19086)
8. Remettre les billes en suspension en vortexant de 30 à 60 sec
9. Placer les billes sur le séparateur magnétique et jeter le surnageant
10. Resuspendre les billes dans le PEG surnageant mis de côté.

Les produits de PCR sont purifiés en deux fractions avec des concentrations de billes de respectivement 1X et 0.4X.

Regrouper 6 x 50µL des puits de PCR et ajouter 1X de billes AMPure XP (fraction 1). Regrouper 10 x 50 µL des puits de PCR et ajouter 0.4X de billes AMPure XP (fraction 2). Poser les tubes sur un agitateur VWR® à 2 000 rpm pendant 10 min. Après une rapide centrifugation placer les tubes sur un séparateur magnétique. Une fois le surnageant clarifié, enlever le en aspirant lentement. Ne pas jeter le surnageant il peut resservir si l'ADN n'est pas récupéré à la fin de la procédure. Laver les billes avec de l'éthanol à 70°C frais. Sans enlever les tubes du séparateur magnétique, ajouter l'éthanol jusqu'à remplir à ras bord les tubes (ex : 1.5 mL pour un tube de 1.5 mL). Après 30 sec retirer l'éthanol. Répéter le lavage à l'éthanol pour un total de 2 lavages. Enlever le plus d'éthanol possible à la pipette et laisser les tubes sécher 30 – 60 sec, bouchon ouvert. Ajouter l'eau sans nucléase 100 µL pour la fraction 1 et 22 µL pour la fraction 2. Laisser incuber 2 min à température ambiante et placer les échantillons sur le séparateur magnétique puis récupérer le surnageant. La fraction 2 peut être conservé à -20°C.

Pour la fraction 1, une seconde purification est réalisée à 1X d'AMPure XP. En répétant le même protocole que celui de la première purification à 1X. L'ADNc est resuspendu à la fin dans 22 µL d'eau sans nucléase.

Puis pour les fractions 1 et 2, faire un dosage Qubit™ (dsDNA BR assay) et un electrophorégramme par bioAnalyzer (Agilent 2200 TapeStation System, Genomic DNA ScreenTape Analysis).

Une fois les concentrations et la taille de librairies identifiées, faire un pool des fractions 1 et 2. Si les concentrations le permettent faire un pool equimolaire avec au moins 1 µg de chaque fraction. Convertir la concentration massique en concentration molaire par l'équation :

$$[ADN]_{molaire,nM} = \frac{[ADN]_{massique,ng/\mu L} \times 10^6}{(660 \times taille\ librairies\ (nt))}$$

5. Capture des librairies

Ajuster les concentrations des librairies pour avoir 1 µg/3.5µL. A partir des trousses SureSelectXT Target Enrichment box 1/box 2, préparer le tampon d'hybridation :

Réactifs	Volume (pour 1 ech)
SureSelect Hyb #1	25 µL
SureSelect Hyb #2	1 µL
SureSelect Hyb #3	10 µL
SureSelect Hyb #4	13 µL
Volume total	**49 µL**

Préparer aussi le mélange de bloqueurs :

Réactifs	Volume (pour 1 ech)
SureSelect Indexing Block #1	2.5 µL
SureSelect Block #2	2.5 µL
SureSelect indexing Indexing Block #3	0.6 µL
Blocker polyT Oligo*	1 µL
Blocker SMARTer PCR Oligo*	1 µL
Volume total	**7.6 µL**

* Ces bloqueurs ont été rajoutés au réactifs initiaux d'Agilent pour empêcher la fixation non spécifique des sondes à la queue polyA des ARNm et à l'adaptateur de SMARTer.

Diluer la RNase Block au 1/10ème, 0.3 µL pour 2.7 µL d'eau sans nucléase pour chaque échantillon.

Pour finir préparer le mélange pour la capture :

Réactifs	Volume (pour 1 ech)
RNase Block diluée	3 µL
Sondes de capture	2 µL
Volume total	**5 µL**

Mélanger les 3.5 µL de librairies avec les 7.6 µL du mélange des bloqueurs. Placer les échantillons sur le thermocycleur et lancer le programme 95°C pendant 5 min et 65°C à l'infini. Placer les tampon d'hybridation à 65°C pendant au moins 5 min. Déposer 5 µL du mélange pour la capture sur le thermocycleur. Laisser incuber 2 min à 65°C. Ajouter les 11 µL des librairies + bloqueurs au 5 µL de capture, puis mettre 13 µL du tampon d'hybridation à 65°C sur les échantillons, mélanger par aspiration-refoulement, pour un volume final d'environ 29 µL.

Laisser incuber ce mélange 24h à 65°C.

Préparer les billes de streptavidine T1 et prechauffer le tampon de lavage SureSelect Wash Buffer 2 à 65°C. Resuspendre les billes de streptavidine, pour chaque échantillon compter 50 µL de billes. Ajouter 200 µL de SureSelect Binding Buffer pour chaque 50 µL de billes. Mélanger et placer les billes sur un séparateur magnétique. Enlever et jeter le surnageant. Répéter ces étapes pour un total de 3 lavages. A la fin resuspendre les billes dans 200 µL de SureSelect Binding Buffer.

Après 24h d'incubation, déposer 29 µL du mélange librairies + capture sur les 200 µL de billes de stréptavidine. Incuber à température ambiante pendant 30 min sur un agitateur. Puis centrifuger brièvement les échantillons et placer les sur un séparateur magnétique ; enlever et jeter le surnageant. Resuspendre les billes dans 200 µL de SureSelect Wash Buffer 1. Incuber 15 min à température ambiante. Remettre les billes sur le séparateur magnétique et jeter le surnageant. Resuspendre les billes dans 200 µL de SureSelect Wash Buffer 2 préchauffé à 65°C. Incuber 10 min à 65°C, mélanger par

retournement au bout de 5 min. centrifuger brièvement et placer les échantillons sur le séparateur magnétique. Enlever et jeter le surnageant. Répeter ces étapes pour un total de 3 lavages.

A la fin des lavages, resuspendre les billes dans 50 µL de SureSelect Elution Buffer, incuber 10 min à température ambiante. Placer les échantillons sur le séparatuer magnétique et récupérer le surnageant. Aux 50 µL d'échantillons ajouter 50 µL de SureSelect Neutralization Buffer. Lavage des échantillons par les billes AMPure XP 1X pour récupérer les échantillons dans un volume final de 50 µL d'eau sans nucléase. Pour cela après avoir ajoutées les billes et laisser incuber 5 min à température ambiante. Après une rapide centrifugation placer les tubes sur un séparateur magnétique. Une fois le surnageant clarifié, enlever le en aspirant lentement. Laver les billes avec de l'éthanol à 70°C frais. Sans enlever les tubes du séparateur magnétique, ajouter 200µL d'éthanol. Après 30 sec retirer l'éthanol. Répéter le lavage à l'éthanol pour un total de 2 lavages. Enlever le plus d'éthanol possible à la pipette et laisser les tubes sécher 30 – 60 sec, bouchon ouvert. Ajouter 50 µL d'eau sans nucléase. Laisser incuber 2 min à température ambiante et placer les échantillons sur le séparateur magnétique puis récupérer le surnageant. Les produits de capture peuvent être stockés à -20°C.

6. PCR post-capture

Avec la trousse de réactif de Takara LA Taq DNA, préparer le mélange réactionnel suivant :

Réactifs	**Volume (pour 1 ech)**
Eau sans nucléase	104.5 µL
10x LA PCR Buffer	20 µL
dNTP (2.5 mM chacun)	16 µL
SMARTer PCR Oligos (12 µM)	8.3 µL
Takara LA Taq DNA polymerase	1.2 µL
Echantillon	50 µL
Volume total	**200 µL**

Diviser en deux ce mélange pour avoir 100 µL de volume. Puis procéder à la PCR avec le programme :

Dénaturation initiale :

- 95°C pour 2 min

14 cycles :

- 95°C pour 20 sec
- 68°C pour 10 min

Extension finale :

- 72°C pour 10 min

Après la PCR regrouper les deux aliquots d'échantillon pour avoir 200 µL d'échantillon. Puis lavage avec les billes AMPure 1X. Pour cela après avoir ajoutées les billes et laisser incuber 5 min à température

ambiante. Après une rapide centrifugation placer les tubes sur un séparateur magnétique. Une fois le surnageant clarifié, enlever le en aspirant lentement. Laver les billes avec de l'éthanol à 70°C frais. Sans enlever les tubes du séparateur magnétique, ajouter 200µL d'éthanol. Après 30 sec retirer l'éthanol. Répéter le lavage à l'éthanol pour un total de 2 lavages. Enlever le plus d'éthanol possible à la pipette et laisser les tubes sécher 30 – 60 sec, bouchon ouvert. Ajouter 50 µL d'eau sans nucléase. Laisser incuber 2 min à température ambiante et placer les échantillons sur le séparateur magnétique puis récupérer le surnageant. Faire un dosage Qubit™ (dsDNA BR assay) et un electrophorégramme par bioAnalyzer (Agilent 2200 TapeStation System, Genomic DNA ScreenTape Analysis). Les produits de PCR peuvent être stockés à -20°C.

7. Librairies construction

Utiliser et suivre le protocole de SMRTbell Library Construction de PacBio.

8. Analyses bioinformatiques

Deux analyses bioinformatiques ont été faites en parallèles. La première avec le pipeline Iso-Seq de PacBio. La seconde par nos propres outils bioinformatiques. L'analyse par Iso-Seq a été réalisé par la plateforme SiRIC. Pour la seconde analyse nous avons utilisés l'outil STAR v2.6.0 pour l'alignement sur le génome hg19 et le comptage de *read* par HTSeq count v0.6.1. Les options définies pour STAR étaient :

```
/Path/to/STARlong \
     --outSAMstrandField intronMotif \
     --outFilterMismatchNmax 2 \
     --outFilterMultimapNmax 10 \
     --genomeDir /Path/to/genome/ \
     --readFilesIn /Path/to/file.fastq \
     --runThreadN 5 \
     --outSAMunmapped Within \
     --seedPerReadNmax 10000 \
     --outSAMtype BAM SortedByCoordinate \
     --limitBAMsortRAM 15000000000 \
     --outSAMheaderHD @HD VN:1.4 SO:SortedByCoordinate \
     --outFileNamePrefix ./file.fastq. \
     --genomeLoad LoadAndKeep
```

Pour permettre la visualisation des différentes isoformes, les fichiers BAM sont convertis en fichier BED par l'outil BEDtools v2.17, avec la fonction bamToBed.

VII. REFERENCES ANNEXES

[1] G.-S. Wang and T. A. Cooper, 'Splicing in disease: disruption of the splicing code and the decoding machinery', *Nat. Rev. Genet.*, vol. 8, no. 10, pp. 749–761, Oct. 2007.

[2] N. López-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigó, 'Are splicing mutations the most frequent cause of hereditary disease?', *FEBS Lett.*, vol. 579, no. 9, pp. 1900–1903, 2005.

[3] R. Cheung *et al.*, 'A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions', *Mol. Cell*, vol. 73, no. 1, pp. 183-194.e8, Jan. 2019.

[4] K. Wimmer *et al.*, 'Extensive in silico analysis of NF1 splicing defects uncovers determinants for splicing outcome upon 5′ splice-site disruption', *Hum. Mutat.*, vol. 28, no. 6, pp. 599–612, 2007.

[5] C. Houdayer *et al.*, 'Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants', *Hum. Mutat.*, vol. 33, no. 8, pp. 1228–1238, Aug. 2012.

[6] X. Jian, E. Boerwinkle, and X. Liu, 'In silico prediction of splice-altering single nucleotide variants in the human genome', *Nucleic Acids Res.*, vol. 42, no. 22, pp. 13534–13544, Dec. 2014.

[7] R. Leman *et al.*, 'Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort', *Nucleic Acids Res.*, vol. 46, no. 21, pp. 11656–11657, Nov. 2018.

[8] R. Leman *et al.*, 'Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants', *BMC Genomics*, In revision.

[9] O. Soukarieh *et al.*, 'Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools', *PLOS Genet.*, vol. 12, no. 1, p. e1005756, Jan. 2016.

[10] H. Tubeuf *et al.*, 'Recommendations to prioritize genetic variants for RNA analyses derived from a large-scale performance evaluation of splicing regulation-predictors', In revision.

[11] L. Grodecká, E. Buratti, and T. Freiberger, 'Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics?', *Int. J. Mol. Sci.*, vol. 18, no. 8, p. 1668, Aug. 2017.

[12] D. R. Zerbino *et al.*, 'Ensembl 2018', *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, Jan. 2018.

[13] J. Zuallaert, F. Godin, M. Kim, A. Soete, Y. Saeys, and W. De Neve, 'SpliceRover: interpretable convolutional neural networks for improved splice site prediction', *Bioinformatics*, vol. 34, no. 24, pp. 4180–4188, Dec. 2018.

[14] H. Y. Xiong *et al.*, 'The human splicing code reveals new insights into the genetic determinants of disease', *Science*, vol. 347, no. 6218, p. 1254806, Jan. 2015.

[15] K. Jaganathan *et al.*, 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, vol. 176, no. 3, pp. 535-548.e24, Jan. 2019.

[16] The 1000 Genomes Project Consortium, 'A global reference for human genetic variation', *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015.

[17] W. J. Kent *et al.*, 'The Human Genome Browser at UCSC', *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jan. 2002.

[18] M. J. Landrum *et al.*, 'ClinVar: public archive of interpretations of clinically relevant variants', *Nucleic Acids Res.*, vol. 44, no. Database issue, pp. D862–D868, Jan. 2016.

[19] W. McLaren *et al.*, 'The Ensembl Variant Effect Predictor', *Genome Biol.*, vol. 17, no. 1, p. 122, Jun. 2016.

[20] S. Monger, M. Troup, E. Ip, S. L. Dunwoodie, and E. Giannoulatou, 'Spliceogen: An integrative, scalable tool for the discovery of splice-altering variants.', *Bioinforma. Oxf. Engl.*, Apr. 2019.

[21] A. Anna and G. Monika, 'Splicing mutations in human genetic disorders: examples, detection, and confirmation', *J. Appl. Genet.*, vol. 59, no. 3, pp. 253–268, Aug. 2018.

[22] M. A. Lewandowska, 'The missing puzzle piece: splicing mutations', *Int. J. Clin. Exp. Pathol.*, vol. 6, no. 12, pp. 2675–2682, Nov. 2013.

[23] A. Woolfe, J. C. Mullikin, and L. Elnitski, 'Genomic features defining exonic variants that modulate splicing', *Genome Biol.*, vol. 11, no. 2, p. R20, Feb. 2010.

[24] E. Buratti *et al.*, 'Aberrant 5′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization', *Nucleic Acids Res.*, vol. 35, no. 13, pp. 4250–4263, Jul. 2007.

[25] I. Vořechovský, 'Aberrant 3′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization', *Nucleic Acids Res.*, vol. 34, no. 16, pp. 4630–4641, Sep. 2006.

[26] A. B. Spurdle *et al.*, 'ENIGMA—Evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes', *Hum. Mutat.*, vol. 33, no. 1, pp. 2–7, 2012.
[27] F. Riant, M. Cecillon, P. Saugier-Veber, and E. Tournier-Lasserve, 'CCM molecular screening in a diagnosis context: novel unclassified variants leading to abnormal splicing and importance of large deletions', *neurogenetics*, vol. 14, no. 2, pp. 133–141, May 2013.
[28] A. Sabbagh *et al.*, 'NF1 Molecular Characterization and Neurofibromatosis Type I Genotype–Phenotype Correlation: The French Experience', *Hum. Mutat.*, vol. 34, no. 11, pp. 1510–1518, 2013.
[29] I. Callebaut *et al.*, 'Comprehensive functional annotation of 18 missense mutations found in suspected hemochromatosis type 4 patients', *Hum. Mol. Genet.*, vol. 23, no. 17, pp. 4479–4490, Sep. 2014.
[30] P. Gaildrat, A. Killian, A. Martins, I. Tournier, T. Frébourg, and M. Tosi, 'Use of Splicing Reporter Minigene Assay to Evaluate the Effect on Splicing of Unclassified Genetic Variants', in *Cancer Susceptibility: Methods and Protocols*, M. Webb, Ed. Totowa, NJ: Humana Press, 2010, pp. 249–257.
[31] A. Y. Steffensen *et al.*, 'Functional characterization of *BRCA1* gene variants by mini-gene splicing assay', *Eur. J. Hum. Genet.*, vol. 22, no. 12, pp. 1362–1368, Dec. 2014.
[32] M. B. Shapiro and P. Senapathy, 'RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression.', *Nucleic Acids Res.*, vol. 15, no. 17, pp. 7155–7174, Sep. 1987.
[33] G. Yeo and C. B. Burge, 'Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals', *J. Comput. Biol.*, vol. 11, no. 2–3, pp. 377–394, Mar. 2004.
[34] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud, 'Human Splicing Finder: an online bioinformatics tool to predict splicing signals', *Nucleic Acids Res.*, vol. 37, no. 9, pp. e67–e67, May 2009.
[35] T. R. Mercer *et al.*, 'Genome-wide discovery of human splicing branchpoints', *Genome Res.*, p. gr.182899.114, Jan. 2015.
[36] A. Corvelo, M. Hallegger, C. W. J. Smith, and E. Eyras, 'Genome-Wide Association between Branch Point Properties and Alternative Splicing', *PLOS Comput. Biol.*, vol. 6, no. 11, p. e1001016, Nov. 2010.
[37] Q. Zhang *et al.*, 'BPP: a sequence-based algorithm for branch point prediction', *Bioinformatics*, vol. 33, no. 20, pp. 3166–3172, Oct. 2017.
[38] B. Signal, B. S. Gloss, M. E. Dinger, T. R. Mercer, and J. Hancock, 'Machine learning annotation of human branchpoints', *Bioinformatics*, vol. 34, no. 6, pp. 920–927, Mar. 2018.
[39] J. M. Paggi and G. Bejerano, 'A sequence-based, deep learning model accurately predicts RNA splicing branchpoints', *RNA*, p. rna.066290.118, Sep. 2018.
[40] I. Nazari, H. Tayara, and K. T. Chong, 'Branch Point Selection in RNA Splicing Using Deep Learning', *IEEE Access*, vol. 7, pp. 1800–1807, 2019.
[41] D. D. Giacomo *et al.*, 'Functional Analysis of a Large set of BRCA2 exon 7 Variants Highlights the Predictive Value of Hexamer Scores in Detecting Alterations of Exonic Splicing Regulatory Elements', *Hum. Mutat.*, vol. 34, no. 11, pp. 1547–1557, 2013.
[42] S. Ke *et al.*, 'Quantitative evaluation of all hexamers as exonic splicing elements', *Genome Res.*, vol. 21, no. 8, pp. 1360–1374, Jan. 2011.
[43] S. N. Teraoka *et al.*, 'Splicing Defects in the Ataxia-Telangiectasia Gene, ATM: Underlying Mutations and Consequences', *Am. J. Hum. Genet.*, vol. 64, no. 6, pp. 1617–1631, Jun. 1999.
[44] I. Tournier *et al.*, 'A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects', *Hum. Mutat.*, vol. 29, no. 12, pp. 1412–1424, 2008.
[45] K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, and W. G. Fairbrother, 'Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes', *Proc. Natl. Acad. Sci.*, vol. 108, no. 27, pp. 11093–11098, Jul. 2011.
[46] T. Sterne-Weiler, J. Howard, M. Mort, D. N. Cooper, and J. R. Sanford, 'Loss of exon identity is a common mechanism of human inherited disease', *Genome Res.*, vol. 21, no. 10, pp. 1563–1571, Jan. 2011.
[47] W. F. Mueller, L. S. Z. Larsen, A. Garibaldi, G. W. Hatfield, and K. J. Hertel, 'The Silent Sway of Splicing by Synonymous Substitutions', *J. Biol. Chem.*, vol. 290, no. 46, pp. 27700–27711, Nov. 2015.
[48] R. Soemedi *et al.*, 'Pathogenic variants that alter protein code often disrupt splicing', *Nat. Genet.*, vol. 49, no. 6, pp. 848–855, Jun. 2017.
[49] I. Lopez-Perolio *et al.*, 'Alternative splicing and ACMG-AMP-2015-based classification of PALB2 genetic variants: an ENIGMA report', *J. Med. Genet.*, p. jmedgenet-2018-105834, Mar. 2019.
[50] G. Davy *et al.*, 'Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer', *Eur. J. Hum. Genet.*, vol. 25, no. 10, pp. 1147–1154, Oct. 2017.
[51] R. Chaligné *et al.*, 'The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer', *Genome Res.*, vol. 25, no. 4, pp. 488–503, Jan. 2015.

[52] M. Colombo *et al.*, 'Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium', *Hum. Mol. Genet.*, vol. 23, no. 14, pp. 3666–3680, Jul. 2014.
[53] L. H. Saal *et al.*, 'The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine', *Genome Med.*, vol. 7, no. 1, Feb. 2015.
[54] A. Dobin *et al.*, 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.
[55] P. J. Whiley *et al.*, 'Comparison of mRNA Splicing Assay Protocols across Multiple Laboratories: Recommendations for Best Practice in Standardized Clinical Testing', *Clin. Chem.*, vol. 60, no. 2, pp. 341–352, Feb. 2014.
[56] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, 'TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biol.*, vol. 14, no. 4, p. R36, Apr. 2013.
[57] S. Farber-Katz *et al.*, 'Quantitative Analysis of BRCA1 and BRCA2 Germline Splicing Variants Using a Novel RNA-Massively Parallel Sequencing Assay', *Front. Oncol.*, vol. 8, 2018.
[58] S. Schafer, K. Miao, C. C. Benson, M. Heinig, S. A. Cook, and N. Hubner, 'Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI)', *Curr. Protoc. Hum. Genet.*, vol. 87, no. 1, pp. 11.16.1-11.16.14, 2015.
[59] B. Rodríguez-Martín *et al.*, 'ChimPipe: accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data', *BMC Genomics*, vol. 18, no. 1, p. 7, Jan. 2017.
[60] J. Pauty *et al.*, 'Cancer-causing mutations in the tumor suppressor PALB2 reveal a novel cancer mechanism using a hidden nuclear export signal in the WD40 repeat motif', *Nucleic Acids Res.*, vol. 45, no. 5, pp. 2644–2657, Mar. 2017.
[61] T. C. Nepomuceno, G. De Gregoriis, F. M. B. De Oliveira, G. Suarez-Kurtz, A. N. Monteiro, and M. A. Carvalho, 'The Role of PALB2 in the DNA Damage Response and Cancer Predisposition', *Int. J. Mol. Sci.*, vol. 18, no. 9, p. 1886, Sep. 2017.
[62] J.-Y. Park, F. Zhang, and P. R. Andreassen, 'PALB2: The hub of a network of tumor suppressors involved in DNA damage responses', *Biochim. Biophys. Acta BBA - Rev. Cancer*, vol. 1846, no. 1, pp. 263–275, Aug. 2014.
[63] A. W. Oliver, S. Swift, C. J. Lord, A. Ashworth, and L. H. Pearl, 'Structural basis for recruitment of BRCA2 by PALB2', *EMBO Rep.*, vol. 10, no. 9, pp. 990–996, Sep. 2009.
[64] S. Reid *et al.*, 'Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer', *Nat. Genet.*, vol. 39, no. 2, pp. 162–164, Feb. 2007.
[65] S. M. H. Sy, M. S. Y. Huen, and J. Chen, 'PALB2 is an integral component of the BRCA complex required for homologous recombination repair', *Proc. Natl. Acad. Sci.*, vol. 106, no. 17, pp. 7155–7160, Apr. 2009.
[66] T. K. Foo *et al.*, 'Compromised BRCA1–PALB2 interaction is associated with breast cancer risk', *Oncogene*, vol. 36, no. 29, pp. 4161–4170, Jul. 2017.
[67] J.-Y. Bleuyard, R. M. Butler, and F. Esashi, 'Perturbation of PALB2 function by the T413S mutation found in small cell lung cancer', *Wellcome Open Res.*, vol. 2, Jan. 2018.
[68] J.-Y. Park *et al.*, 'Breast cancer-associated missense mutants of the PALB2 WD40 domain, which directly binds RAD51C, RAD51 and BRCA2, disrupt DNA repair', *Oncogene*, vol. 33, no. 40, pp. 4803–4812, Oct. 2014.
[69] P. J. Byrd *et al.*, 'A Hypomorphic PALB2 Allele Gives Rise to an Unusual Form of FA-N Associated with Lymphoid Tumour Development', *PLOS Genet.*, vol. 12, no. 3, p. e1005945, Mar. 2016.
[70] C. Xu and J. Min, 'Structure and function of WD40 domain proteins', *Protein Cell*, vol. 2, no. 3, pp. 202–214, Mar. 2011.
[71] M. de la Hoya *et al.*, 'Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms', *Hum. Mol. Genet.*, vol. 25, no. 11, pp. 2256–2268, Jun. 2016.
[72] A. Acedo, C. Hernández-Moro, Á. Curiel-García, B. Díez-Gómez, and E. A. Velasco, 'Functional Classification of BRCA2 DNA Variants by Splicing Assays in a Large Minigene with 9 Exons', *Hum. Mutat.*, vol. 36, no. 2, pp. 210–221, 2015.
[73] E. Fraile-Bethencourt, B. Díez-Gómez, V. Velásquez-Zapata, A. Acedo, D. J. Sanz, and E. A. Velasco, 'Functional classification of DNA variants by hybrid minigenes: Identification of 30 spliceogenic variants of BRCA2 exons 17 and 18', *PLOS Genet.*, vol. 13, no. 3, p. e1006691, Mar. 2017.
[74] A. V. Bryksin and I. Matsumura, 'Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids', *BioTechniques*, vol. 48, no. 6, pp. 463–465, Jun. 2010.
[75] P. Gaildrat *et al.*, 'The BRCA1 c.5434C→G (p.Pro1812Ala) variant induces a deleterious exon 23 skipping by affecting exonic splicing regulatory elements', *J. Med. Genet.*, vol. 47, no. 6, pp. 398–403, Jun. 2010.

[76] I. Catucci *et al.*, '*PALB2* sequencing in Italian familial breast cancer cases reveals a high-risk mutation recurrent in the province of Bergamo', *Genet. Med.*, vol. 16, no. 9, pp. 688–694, Sep. 2014.
[77] S. Casadei *et al.*, 'Contribution of Inherited Mutations in the BRCA2-Interacting Protein PALB2 to Familial Breast Cancer', *Cancer Res.*, vol. 71, no. 6, pp. 2222–2229, Mar. 2011.
[78] L. C. Walker *et al.*, 'Evaluation of a 5-Tier Scheme Proposed for Classification of Sequence Variants Using Bioinformatic and Splicing Assay Data: Inter-Reviewer Variability and Promotion of Minimum Reporting Guidelines', *Hum. Mutat.*, vol. 34, no. 10, pp. 1424–1431, Oct. 2013.
[79] J. D. Fackenthal *et al.*, 'Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples', *J. Med. Genet.*, vol. 53, no. 8, pp. 548–558, Aug. 2016.
[80] O. Anczuków *et al.*, 'BRCA2 Deep Intronic Mutation Causing Activation of a Cryptic Exon: Opening toward a New Preventive Therapeutic Strategy', *Clin. Cancer Res.*, vol. 18, no. 18, pp. 4903–4909, Sep. 2012.
[81] J. Dutil *et al.*, 'No Evidence for the Pathogenicity of the BRCA2 c.6937 + 594T>G Deep Intronic Variant: A Case–Control Analysis', *Genet. Test. Mol. Biomark.*, vol. 22, no. 2, pp. 85–89, Jan. 2018.

Développement d'outils biostatistiques et bioinformatiques de prédiction et d'analyse des défauts de l'épissage : application aux gènes de prédisposition aux cancers du sein et de l'ovaire.

Mots-clés : épissage, variants, syndrome HBOC, prédiction, RNA-seq, SPiP, SPiCE, SpliceLauncher

Résumé

L'analyse des défauts d'épissage est particulièrement complexe. Outre la diversité des transcrits présents à l'état physiologique, les variations nucléotidiques peuvent induire des modifications hétéroclites de l'épissage. Ces variations, appelées variants splicéogéniques, et leur impact au niveau de l'épissage, sont à même de modifier plus ou moins sévèrement le phénotype de l'individu.

Au cours de ce travail de thèse, nous nous sommes intéressés à trois grands aspects de l'étude des défauts de l'épissage : (i) la prédiction de ces défauts d'épissage, (ii) l'analyse des données de RNA-seq et (iii) le rôle de l'épissage dans l'interprétation de la pathogénicité d'un variant pour la prédisposition aux cancers du sein et de l'ovaire (syndrome HBOC).

Nous avons optimisé les recommandations en vigueur pour identifier les variants splicéogéniques au sein des séquences consensus des sites d'épissage. Ce travail a conduit à la publication d'un nouvel outil SPiCE (*Splicing Prediction in Consensus Elements*), développé sur 395 variants. SPiCE a le potentiel d'être une aide à la décision pour guider les généticiens vers ces variants splicéogéniques, grâce à une exactitude de 94.4 %. Puis, nous avons comparé les outils de prédiction des points de branchement. Pour cela, une collection sans précédente de 120 variants avec leurs études ARN a été établi dans la région des points de branchements. Nous avons ainsi révélé que ces outils de prédictions sont aptes à prioriser les variants pour des études ARN dans ces régions jusque-là peu étudiées. Pour étendre les prédictions des variants splicéogéniques au-delà d'un motif spécifique, nous avons construit l'outil SPiP (*Splicing Prediction Pipeline*). SPiP utilise un ensemble d'outils pour prédire un défaut d'épissage quel que soit la position du variant. Ainsi, SPiP peut ainsi s'adresser à la diversité des défauts d'épissage avec une exactitude de 80.21 %, sur une collection de 2 784 variants.

Les données issues du RNA-seq sont complexes à analyser, car il existe peu d'outils pour annoter finement les épissages alternatifs. Aussi nous avons publié l'outil SpliceLauncher. Cet outil permet de déterminer une grande diversité de jonctions d'épissage, indépendamment des systèmes RNA-seq utilisés. Cet outil renvoie aussi les résultats sous formes graphiques pour faciliter leur interprétation.

Puis nous avons évalué le rôle de l'épissage alternative dans l'interprétation à usage clinique d'un variant. Le gène *PALB2*, impliqué dans le syndrome HBOC, a été utilisé comme modèle d'étude. Nous avons ainsi démontré que l'épissage alternatif de *PALB2* est apte à remettre en cause la pathogénicité de certains variants. La collecte de données fonctionnelles et cliniques sont donc nécessaires pour conclure sur leur pathogénicité.

Nos travaux illustrent ainsi l'importance de la caractérisation et de l'interprétation des modifications de l'épissage pour répondre aux défis présents et futurs du diagnostic moléculaire en génétique.

Abstract

Analysis of splicing defects is particularly complex. In addition to the diversity of physiological transcripts, nucleotidic variations can induce heterogeneous alteration of splicing. These variations, called spliceogenic variants, and their impact on splicing, can involve severe consequences on the individual phenotype.

In this thesis work, we focused on three main aspects of the study of splicing defects: (i) the prediction of these splicing defects, (ii) the analysis of RNA-seq data and (iii) the role of splicing in interpreting the pathogenicity of a variant for the hereditary breast and ovarian cancers (HBOC syndrome).

We optimized the current recommendations to identify spliceogenic variants within the consensus sequences of splicing sites. This work led to the publication of a new tool, SPiCE (Splicing Prediction in Consensus Elements), developed on 395 variants. SPiCE has the potential to be a decision support tool to guide geneticists towards these spliceogenic variants, with an accuracy of 94.4%. Then, we compared the tools dedicated to branch points prediction. For this purpose, an unprecedented collection of 120 variants with their RNA studies has been established in the branch point region. Thus, we revealed these prediction tools are able to prioritize variants for RNA studies in these hitherto poorly studied regions. To extend the predictions of spliceogenic variants beyond a specific motif, we built SPiP (Splicing Prediction Pipeline) tool. SPiP uses a set of tools to predict a splicing defect regardless of the variant position. Thus, SPiP can address the diversity of splicing defects with an accuracy of 80.21%, on a collection of 2,784 variants.

The data from the RNA-seq are complex to analyze, as there are few tools to finely annotate alternative splices. Also we published SpliceLauncher tool. This tool allows to determine a wide variety of splicing junctions, independently of RNA-seq systems used. This tool also returns the results in graphical form to make interpretation user-friendly.

Then we evaluated the role of alternative splicing in the clinical interpretation of a variant. The *PALB2* gene, involved in HBOC syndrome, was used as a study model. Thus, we demonstrated that the alternative splicing of *PALB2* is able of challenging the pathogenicity of certain variants. Collection of functional and clinical data is therefore necessary to conclude on their pathogenicity.

Our work thus illustrates the importance of characterizing and interpreting splicing modifications to meet the current and future challenges of molecular diagnosis in human genetics.