Développement de stratégies de criblage de mutations d'épissage dans des gènes de prédisposition au cancer.

Helene Tubeuf

Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité Aspects moléculaires et cellulaires de la biologie
– Génétique du cancer –

Préparée au sein de l'Université de Rouen Normandie

Développement de stratégies de criblage de mutations d'épissage dans les gènes de prédispostions au cancer

Présentée et soutenue par
Hélène TUBEUF

Thèse soutenue publiquement le 16 Avril 2019 devant le jury composé de		
Mme Dominique STOPPA-LYONNET	PR, Institut Curie	Rapporteur
Mme Sylvie TUFFERY-GIRAUD	IR, Université de Montpellier	Rapporteur
Mme Pascale FANEN	PU-PH, Université de Paris-Est Créteil	Examinateur
M. Laurent CORCOS	DR, Université de Bretagne Occidentale	Examinateur
M. Thierry FREBOURG	PU-PH, Université de Rouen Normandie	Examinateur
M. André BLAVIER	Directeur Scientifique, Interactive Biosoftware	Membre invité
Mme Alexandra MARTINS	CR1, Université de Rouen Normandie	Directrice de thèse

Thèse dirigée par Dr. Alexandra MARTINS
Unité Inserm U1245 - Génomique et médecine personnalisée du cancer et des maladies neuropsychiatriques

UNIVERSITÉ DE ROUEN NORMANDIE

EdNBISE
Ecole doctorale Normande
Biologie Intégrative,
Santé, Environnement

Inserm
La science pour la santé
From science to health

RESUME

Le développement du séquençage de l'ADN à haut débit a grandement facilité le criblage de variations génétiques dans le génome des patients. Désormais, l'un des principaux défis de la génétique médicale n'est donc plus la détection des variations, mais leur interprétation fonctionnelle et clinique. Récemment, nous avons montré, à l'aide de tests fonctionnels basés sur l'utilisation de minigènes, que bien que le nombre de mutations d'épissage, et en particulier celles qui affectent sa régulation, est actuellement sous-estimé, l'effet de ces variations pourrait être dorénavant prédit à l'aide d'outils bioinformatiques spécifiques. Nous avons ainsi étendu l'évaluation du caractère prédictif de ces quatre nouvelles approches bioinformatiques par une étude comparative des scores générés par ces approches avec des données expérimentales obtenues pour un total d'environ 1200 variations exoniques. Nos travaux ont ainsi démontré la fiabilité de ces approches, utilisées seules ou en combinaison, et ont permis de proposer des recommandations quant à leur utilisation en tant qu'outils de filtration pour prioriser les variations à analyser dans des tests fonctionnels axés sur l'épissage. Néanmoins, une analyse mutationnelle exhaustive ciblée sur l'exon 7 de *MLH1*, a mis en évidence l'échec apparent de ces approches, pourtant validées par des études menées sur l'exon 7 de *BRCA2*, l'exon 10 de *MAPT* et l'exon 5 de *MSH2*, laissant suggérer que ces méthodes pourraient ne pas s'appliquer de manière équivalente à tous les exons et/ou à tous les gènes. En effet, nous avons montré que cet exon était doté de caractéristiques particulières, i.e. de sites d'épissage remarquablement forts, lui conférant une résistance totale aux mutations de régulation d'épissage et mettant en échec les outils de prédictions. Ces données contribuent à mieux déterminer les limitations de ces outils bioinformatiques tout en contribuant à leur amélioration. En dépit de ces avancées, l'évaluation de la pathogénicité des mutations d'épissage reste complexe, en particulier celles conduisant à des anomalies d'épissage en phase et/ou partielles. En utilisant, comme modèle d'étude, des variations à l'origine du saut partiel de l'exon 3 de *BRCA2*, nos résultats ont révélé que l'activité tumeur-suppressive de BRCA2 tolère une réduction substantielle du niveau d'expression, étant donné qu'un allèle produisant jusqu'à 70% de transcrit codant une protéine déficiente n'est pas nécessairement associé à un risque élevé de développer un cancer. L'ensemble de ces données a d'importantes implications dans le diagnostic moléculaire et la prise en charge des patients et de leurs apparentés, avec un bénéfice direct pour les familles évocatrices d'une prédisposition héréditaire et devrait contribuer à l'interprétation de VSI identifiées par séquençage à haut débit dans toute autre pathologie d'origine génétique.

Mots clés : cancer héréditaire, interprétation des variations génétiques, régulation de l'épissage de l'ARN, prédictions bioinformatiques, analyses fonctionnelles.

ASBTRACT

The development of high-throughput DNA sequencing has greatly facilitated the screening of genetic variations within patient genome. Henceforth, one of the main challenges in medical genetics is no longer the detection of variations, but their functional and clinical interpretation. Recently, we showed by using splicing reporter minigene assays, that although splicing mutations, and in particular those affecting its regulation, are more prevalent than initially estimated, they could now be predicted by using dedicated bioinformatics tools. We thus extended the evaluation of the predictive power of these four newly developed computational approaches by a comparative study of the scores obtained by these approaches with experimental data for a total of about 1200 exonic variations. Our findings have demonstrated the reliability of these approaches, used alone or in combination, and allow to offer recommendations for their use as a filtration tool to prioritize the variations to be analysed as a priority in splicing-dedicated functional assays. Nevertheless, an exhaustive mutational analysis targeting *MLH1* exon 7, has highlighted the apparent failure of these approaches, yet validated by studies focused on *BRCA2* exon 7, *MAPT* exon 10 and *MSH2* exon 5, suggesting that these methods might not be equivalently applicable to all exons and/or genes. Indeed, we have shown that this exon has particular characteristics, i.e. remarkably strong splice sites, conferring it a total resistance to splicing regulation mutations and defeating prediction tools. These findings help to better determine the limitations of these bioinformatics tools while contributing to their improvement. In spite of these advances, the pathogenicity assessment of splicing mutations remains complicated, especially of those leading to in-frame and/or partial splicing anomalies. By using variant-induced partial *BRCA2* exon 3 skipping as a model system, we showed that BRCA2 tumor suppressor function tolerates a substantial reduction in expression level, as BRCA2 allele producing as much as 70% of transcript encoding deficient protein may not necessarily confer high-risk of developing cancer. Altogether, these data have important implications in the molecular diagnosis and clinical management of patients and their relatives, with a direct benefit for hereditary cancer-suspected families and should contribute to the interpretation of VSI identified by high throughput sequencing in any other genetic disease.

Key words: hereditary cancer, Interpretation of genetic variation, mRNA splicing regulation, bioinformatics predictions, functional analyses.

REMERCIEMENTS

La thèse de doctorat représente un travail s'inscrivant dans la durée, pendant laquelle de nombreuses personnes se retrouvent ainsi de manière fortuite ou non, pour le pire ou le meilleur, entre le doctorant et son doctorat. Ce sont certaines de ces personnes, qui par l'intérêt manifestés à l'égard de mes travaux de recherche ont permis de les faire progresser, que j'aimerais mettre en avant dans ces remerciements.

Mes remerciements vont tout d'abord aux membres de mon jury qui m'ont fait l'honneur de juger ce travail : le Pr Dominique Stoppa-Lyonnet et le Dr Sylvie Tuffery-Giraud pour avoir accepté la lourde tâche de rapporteur, et mes examinateurs ; le Dr Laurent Corcos et le Pr Pascale Fanen, pour leur participation à cette soutenance de thèse. Je vous prie de trouver ici le témoignage de ma profonde reconnaissance.

J'exprime ma profonde reconnaissance au Pr Thierry Frebourg, directeur de l'unité Inserm U1245, pour m'avoir accueillie au sein de son laboratoire et pour l'énergie qu'il emploie à trouver les moyens nécessaires à son équipe afin qu'elle puisse travailler dans des conditions optimales.

Je tiens à exprimer ma gratitude à ma directrice de thèse, le Dr Alexandra Martins, pour m'avoir accueillie au sein de son équipe et guidée au cours de ce travail. Je lui suis reconnaissante pour les nombreuses discussions scientifiques que nous avons eues, les conseils qu'elle m'a donnés et la confiance qu'elle m'a accordée. Je la remercie également pour l'autonomie et la latitude de travail qu'elle m'a laissées : développement technologique, diagnostic et implication dans des nombreux projets de recherche bien que malheureusement nombreux sont ceux qui n'ont pas encore été valorisés.

Je remercie chaleureusement l'ensemble des membres du groupe « ARN et Cancer » pour leur contribution à la réussite de mon travail de thèse : Sophie, Pascaline, Gaïa, Aurélie, Omar, Raphaël, et Marion, ainsi que la multitude de stagiaires de passage. Gaïa et Aurélie, merci infiniment pour votre bienveillance, votre disponibilité et votre patience, pendant ma formation aux rudiments des tests fonctionnels d'épissage. Une grande pensée à Gaïa, mon mentor à la paillasse, dont la bonne humeur communicative m'a grandement manqué tout au long de ma thèse. J'espère que tu t'épanouies pleinement dans ton nouveau rôle : nul doute que tu vas susciter de nouvelles vocations parmi tes jeunes scientifiques en herbe. Omar, merci pour tes conseils, ton attention et ta gentillesse. Je te souhaite le meilleur pour la poursuite de ta carrière. Sophie et Raphaël, merci pour votre bonne humeur, votre confiance et votre disponibilité. J'ai été ravie de travailler avec vous et je vous souhaite beaucoup de réussite pour les projets de recherche à venir.

Je suis aussi reconnaissante à toutes les personnes de l'U1245 pour leur disponibilité, leur sympathie et leur bonne humeur, en particulier celles et ceux qui ont, à un moment donné, contribué significativement à ma formation et mon bien-être. Merci à Nathalie pour son efficacité redoutable dans les tâches administratives et logistiques, à Camille pour sa pédagogie et pour l'aide qu'elle m'a apportée pour la réalisation des analyses statistiques, à Stéphane pour le temps qu'il a accordé à dépanner et résoudre mes problèmes de séquençage, à Anne pour la patience dont elle a su faire preuve en cas de problème en culture cellulaire, à Magalie pour ses précieux conseils en matière de clonage, à Marion, Paul et Emeline, pour m'avoir épaulée sur mes projets pendant leur stage. Je remercie également Anne-Claire, Camille, Claude, Dominique, Françoise, Kévin, Gaël, Gaëlle, Laëtitia, Ludivine, Myriam, Olivier, Pascale, Raphaël, Sébastien, Sherin, Sophie et Stéphanie qui ont toujours été bienveillants à mon égard et prêts à m'apporter leurs expertises respectives en cas de besoin. J'ai également une pensée pour les anciens doctorants devenus docteurs qui étaient présents à mon arrivée et qui depuis, ont vogué vers d'autres horizons : Yasmine, Hafid, Morgane, Omar, Pierre (et Inès), Estelle, et Alexandre. Je les remercie pour tous les bons moments que l'on a partagé et je leur souhaite, à tous, la meilleure des réussites tant sur le plan professionnel que personnel. A Camille, à qui j'adresse tous mes vœux de réussite pour la poursuite de sa thèse : le meilleur est à venir...

J'exprime également toute ma gratitude aux dirigeants d'Interactive Biosoftware, André Blavier, qui me fait l'honneur de participer à ce jury, et Juliette Renault, de m'avoir donné l'opportunité de réaliser une thèse CIFRE au sein de leur entreprise. Je leur exprime ma profonde reconnaissance pour leur confiance, leur engagement, l'intérêt qu'ils ont porté à mes travaux et les efforts qu'ils ont mis en œuvre pour que mon passage dans l'entreprise se déroule dans des conditions optimales tant sur le plan matériel que sur le plan humain. Je tiens également à remercier tous mes collègues, Alexandre, Amandine, Christelle, Daniel, Florian, Gabriella, Hugues, Jean-Maxime, Lauréline, Madina, Maïmouna Séverine, Thomas et Tiphaine pour leur aide, leur soutien, leur gentillesse et leur bonne humeur ainsi que pour tous les bons moments conviviaux, festifs et gustatifs que nous avons partagés. Je leur souhaite à tous beaucoup de succès et de réussite professionnelle dans leur nouvelle aventure au sein de SOPHiA GENETICS ou ailleurs...

Je suis infiniment reconnaissante au Dr Shyam Sharan, à la tête de l'équipe « Génétique de la susceptibilité au cancer » de m'avoir accueillie au sein de son laboratoire quelques mois et d'avoir pris soin de moi. Je le remercie très sincèrement pour sa disponibilité, son attention, son honnêteté et son extrême bienveillance tant au niveau professionnel que personnel et je lui témoigne de toute mon admiration pour la façon dont il mène ses recherches et gère son équipe, avec passion, dynamise, rigueur, dévotion, accessibilité, et humanité. J'associe à ces remerciements l'ensemble des membres de son équipe, Terry, Eileen, Susan, April, Iben, Stacey, Betty, Linda, Kajal, Suhas et Arun. Je les remercie pour leur accueil, leur gentillesse, leur disponibilité, leur expérience, pour l'aide qu'ils m'ont apporté pour mener à bien mon projet et pour leur savoir qu'ils m'ont transmis sans concession. A Heather, qui m'a accueillie chez elle pendant ces quelques mois et à Terry, qui a fait bien plus que

me prendre en charge au laboratoire, je tiens à leur témoigner toute mon affection et à les remercier pour leur amitié, leur profonde gentillesse, leur bonne humeur, leur ouverture d'esprit et leur altruisme. Merci de m'avoir fait partager la vie quotidienne d'une famille américaine, rencontrer de merveilleuses personnes, découvrir certains des magnifiques endroits que renferme ce pays, connaître les coutumes et des traditions et d'avoir su donner et recevoir. Cette expérience restera l'un de mes plus beaux souvenirs.

Ce travail n'aurait pu être mené à bien sans le soutien de l'Association Nationale de la Recherche et de la Technologie (ANRT), l'*OpenHealth Insititute*, le Cancéropôle Nord-Ouest, l'*European Molecular Biology Organisation* (EMBO), l'Ecole Doctorale Normande de Biologie Intégrative, Santé et Environnement (EdN BISE 497), qui m'ont permis, grâce à une allocation de recherche et diverses aides financières, de me consacrer sereinement à l'élaboration de ma thèse ; ni sans la participation des patients et de leur famille et des collaborateurs qui m'ont permis de bénéficier d'échantillons biologiques et/ou d'informations génétiques/cliniques/familiales, en espérant que les données générées puissent leur bénéficier.

J'en profite pour remercier plus particulièrement Ludivine, Nathalie, Laëtitia, Stéphane, Olivier, Anne, Sophie, Marion, mes copains de labo, si différents et tellement attachants, pour la bonne ambiance qu'ils génèrent quotidiennement et le soutien qu'ils m'ont apporté, surtout ces derniers mois, dans les bons comme dans les mauvais moments. Ces nombreuses heures passées à leur côté ont fait naître de réels liens d'amitiés, qui, j'espère, perdureront. A Nathalie, secrétaire hors pair et multi-tâche, sans qui rien ne fonctionnerait correctement au sein de l'unité, je lui suis infiniment reconnaissante pour toute l'aide qu'elle m'a apporté quotidiennement. A Ludivine, camarade d'études, de travail, de congrès, de sorties et de vacances, devenue une amie inestimable au fil des années, je la remercie sincèrement pour son amitié et ce qu'elle m'apporte.

Au terme de ce parcours, ma reconnaissance indéfectible va enfin à celles et ceux qui me sont chers, et qui bien que quelque peu délaissés ces derniers mois pour achever cette thèse, m'ont assuré de leur soutien affectif, de leurs attentions et de leurs encouragements tout au long de ces 8 dernières années : ma famille (qui ne cesse de s'agrandir) et mes amis (qui sont de plus en plus nombreux). Leur amour, leur confiance et leur soutien sans faille m'auront permis de surmonter bien des épreuves. Je leur dédie ce travail qui est autant le mien que le leur. Enfin, j'ai une pensée toute particulière pour mes grands-parents, Jacques et Raymonde, qui auraient dû être parmi nous...

TABLE DES MATIERES

LISTE DES ABREVIATIONS

ADN : acide désoxyribonucléique

ARN : acide ribonucléoprotéique

ASO : *antisense oligonucleotide*

BER : *base excision repair*

BP : *branch point*

CCR : cancer colorectal

DSB : *double strand break*

ESE : *exonic splicing enhancer*

ESR : *exonic splicing regulatory element*

ESS : *exonic splicing silencer*

FAP : *familial adenomatous polyposis*

HBOC : *hereditary breast and ovarian cancer*

hnRNP : *heterogeneous nuclear ribonucleoprotein particle*

HNPCC : *hereditary nonpolyposis colorectal cancer*

RH : recombination homologue

ICL : *instrastrand croolinks*

IHC : immunohistochimie

ISE : *intronic splicing enhancer*

ISR : *intronic splicing regulatory element*

ISS : *intronic splicing silencer*

MMR : *mismatch repair*

MSI : *microsatellite instability*

MSS : *microsatellite stability*

NHEJ : *non homologous end joining*

NGS : *next generation sequencing*

NMD : *nonsense mediated decay*

ORF : *open reading frame*

PARP : poly-(ADP-riboses) polymérases

PPT : *polypyrimidine tract*

PTC : *premature termination codon*

PTM : *pre-trans-splicing molecule*

RBPs : *RNA binding proteins*

RNP : ribonucléoprotéines

RUST : *regulated unproductive splicing and translation*

SMA : atrophie musculaire spinale

SMaRT *: spliceosomal-mediated RNA trans-splicing*

SLA : sclérose latérale amyotrophique

SNV : *single nucleotide variantion*

SNP : *single nucleotide polymorphism*

snRNA : *small nuclear RNA*

snRNP : *small nuclear ribonucleoprotein*

SPLM : *splice modulators*

SR : serine/arginine

SRE : *splicing regulatory element*

ss : *splice site* (5'/3')

TIL, *tumor infiltrating lymphocytes*

VSI : variation de signification inconnue

UTR : *untranslated region* (5'/3')

LISTE DES PUBLICATIONS

Le Guennec, K., Tubeuf, H., Hannequin, D., Wallon, D., Quenez, O., Rousseau, S., Richard, A.-C., Deleuze, J.-F., Boland, A., Frebourg, T., *et al.* (2018). Biallelic Loss of Function of SORL1 in an Early Onset Alzheimer's Disease Patient. J. Alzheimers Dis. JAD 62, 821–831.

Dominguez-Valentin, M, Evans, DGR, Nakken, S, Tubeuf, H, Vodak, D, Ekstrøm, PO, Nissen, AM, Morak, M, Holinski-Feder, E, Martins, A, et al (2018a) Genetic variants of prospectively demonstrated phenocopies in BRCA1/2 kindreds Hered Cancer Clin Pract 16, 4

Dominguez-Valentin, M, Nakken, S, Tubeuf, H, Vodak, D, Ekstrøm, PO, Nissen, AM, Morak, M, Holinski-Feder, E, Martins, A, Møller, P, et al (2018b) Identification of genetic variants for clinical management of familial colorectal tumors BMC Med Genet 19, 26

Dominguez-Valentin, M, Nakken, S, Tubeuf, H, Vodak, D, Ekstrøm, PO, Nissen, AM, Morak, M, Holinski-Feder, E, Martins, A, Møller, P, et al (2018c) Potentially pathogenic germline CHEK2 c319+2T>A among multiple early-onset cancer families Fam Cancer 17, 141–153

Caputo, S.M., Léone, M., Damiola, F., Ehlen, A., Carreira, A., Gaidrat, P., Martins, A., Brandão, R.D., Peixoto, A., Vega, A., Houdayer C, Delnatte C, Bronner M, Muller D, Castera L, Guillaud-Bataille M, Søkilde I, Uhrhammer N, Demomntety S, Tubeuf H, Castealin G, French COVAR group collaborators, Jensen U, Petitalot A, Krieger S, Lefol C, Moncoutier V, Boutry-Kryza N, Nielsen H, Sinilnikova O, Stoppa-Lyonnet D, Spurdle A, Teixeira M, Coulet F, Thomassen M, Rouleau E. (2018). Full in-frame exon 3 skipping of BRCA2 confers high risk of breast and/or ovarian cancer. Oncotarget 9, 17334–17348.

Rossi BM, Palmero EI, López-Kostner F, Sarroca C, Vaccaro CA, Spirandelli F, Ashton-Prolla P, Rodriguez Y, de Campos Reis Galvão H, Reis RM, Escremim de Paula A, Capochin Romagnolo LG, Alvarez K, Della Valle A, Neffa F, Kalfayan PG, Spirandelli E, Chialina S, Gutiérrez Angulo M, Castro-Mujica MDC, Sanchez de Monte J, Quispe R, da Silva SD, Rossi NT, Barletta-Carrillo C, Revollo S, Taborga X, Morillas LL, Tubeuf H, Monteiro-Santos EM, Piñero TA, Dominguez-Barrera C, Wernhoff P, Martins A, Hovig E, Møller P, Dominguez-Valentin M, *A survey of the clinicopathological and molecular characteristics of patients with suspected Lynch syndrome in Latin America*. BMC Cancer 17, 623 (2017)

Gaildrat P, Lebbah S, Tebani A, Sudrié-Arnaud B, Tostivint I, Bollee G, Tubeuf H, Charles T, Bertholet-Thomas A, Goldenberg A, Barbey F, Martins A, Saugier-Veber P, Frébourg T, Knebelmann B, Bekri S, *Clinical and molecular characterization of cystinuria in a French cohort: relevance of assessing large-scale rearrangements and splicing variants*. Mol Genet Genomic Med 5, 373-389 (2017)

Harding BN, Moccia A, Drunat S, Soukarieh O, Tubeuf H, Chitty LS, Verloes A, Gressens P, El Ghouzzi V, Di Cunto F, Martins A, Passemard S, Bielas SL, *Mutations in Citron Kinase Cause Recessive Microlissencephaly with Multinucleated Neurons*. American Journal Human Genetics 99, 511-520 (2016)

INDEX DES FIGURES

INDEX DES TABLEAUX

Partie I : Introduction Générale

CHAPITRE I : PREDISPOSITIONS HEREDITAIRES AUX CANCERS LES PLUS FREQUENTES

1) Les prédispositions héréditaires au cancer

Même si les cancers sont considérés le plus souvent comme des maladies génétiques acquises de présentation sporadique, environ 5-10% de l'ensemble des cancers correspondent à des maladies génétiques héréditaires se transmettant selon un mode Mendélien (Figure 1 ; pour revue : Garsber and Offit, 2005), ce qui représenterait environ 300 000 nouveaux cas par an dans le monde (pour revue : Rahman, 2014). Il a d'ailleurs été dénombré, à ce jour, pas moins de deux cents syndromes associés à des prédispositions génétiques au cancer (Tableau 1 ; pour revues ; Lynch *et al.*, 2015; Rahman, 2014). Ces prédispositions résultent d'une altération constitutionnelle, souvent de pénétrance incomplète, survenant au niveau d'un gène qualifié de gène de prédisposition au cancer et défini comme un gène dont les altérations constitutionnelles conduisent à une augmentation du risque relatif de développer un ou plusieurs cancers chez le sujet porteur, par rapport au risque moyen observé dans la population générale (Tableau 1 ; pour revue : Nagy *et al.*, 2004). La majorité des prédispositions au cancer connues sont transmises d'une génération à l'autre d'une manière autosomique dominante (Tableau 1 ; pour revue : Nagy *et al.*, 2004).

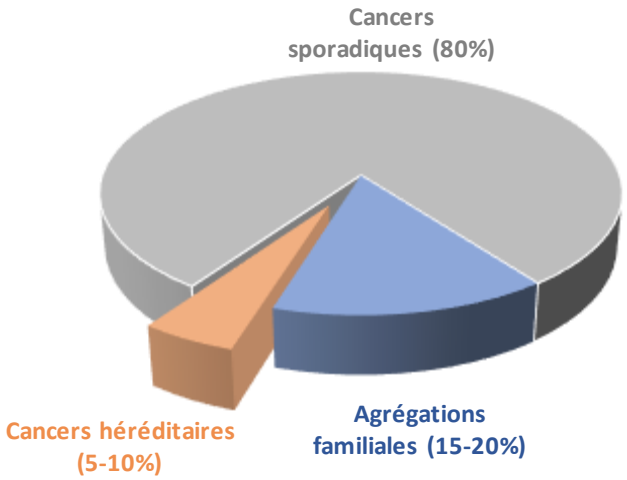


Figure 1 : Fréquence des différents types de cancer (adaptée de Nagy *et al.*, 2004). Bien que la majorité des cancers sont sporadiques (80%), environ 5-10% de l'ensemble des cancers sont héréditaires et résultent d'une mutation constitutionnelle dans un gène de prédisposition au cancer. Environ 15-20% correspondent à des agrégations familiales de cancers et résultent possiblement d'interactions entre des altérations faiblement pénétrantes dans plusieurs gènes, d'interactions avec l'environnement ou les deux.

Tableau 1 : Exemples de syndromes de prédisposition héréditaire au cancer (d'après Garber and Offit, 2005; Nagy *et al.*, 2004; Sifri *et al.*, 2004).

Syndrome	Gène(s) essentiel(s) impliqué(s)	Cancer(s) associé(s)	Incidence dans la population	Pénétrance	Mode de transmission
Syndrome de Lynch (LS)	MLH1 MSH2 MSH6 PMS2	Cancer colorectal Cancer de l'endomètre Cancer de l'ovaire Cancer du pelvis rénal Cancer de l'urètre Cancer du pancréas Cancers de l'estomac et de la vessie Cancers hépatobiliaires	1/400	90%	Dominant
Cancer héréditaire du sein et de l'ovaire (HBOC)	BRCA1 BRCA2	Cancer du sein Cancer de l'ovaire Cancer de la prostate Cancer du pancréas	1/500 - 1/1000	85%	Dominant
Neurofibromatose de type 1	NF1	Neurofibrosarcomes Phéochromocytomes Gliomes optiques Méningiomes	1/3000	100%	Dominant
Polypose adénomateuse familiale (FAP)	APC	Cancer colorectal Hépatoblastomes et cancer de l'intestin grêle Cancer des voies biliaires Cancer de l'estomac Cancer de la thyroïde	1/5000 – 1/10000	~ 100%	Dominant
Neurofibromatose de type 2	NF2	Schwannomes vestibulaires	1/40000	100%	Dominant
Syndrome de Cowden (CS)	PTEN	Cancer du sein Cancer de la thyroïde Cancer de l'endomètre et autres cancers	1/200000	90-95%	Dominant
Anémie de Fanconi (FA)	FANCA FANCB FANCC FANCD FANCE FANCF FANCG FANCL BRCA2	Leucémies Cancers squameux Carcinomes de la peau Hépatomes	1/360000	100%	Récessif
Syndrome de Li-Fraumeni (LFS)	TP53	Sarcomes des tissus mous Cancer du sein Ostéosarcomes Leucémies Tumeurs cérébrales Corticosurrénalomes	Rare	90-95%	Dominant

Les prédispositions héréditaires au cancer présentent, au même titre que les maladies Mendéliennes, un certain nombre de signes fortement évocateurs tels que (i) la survenue d'un cancer à un âge précoce, (ii) une histoire familiale de cancer, avec l'existence de plusieurs cas de cancers sur plusieurs générations, (iii) l'existence de tumeurs primitives multiples chez un individu avec notamment les formes multifocales et bilatérales pour les organes doubles et (iv) une association d'une pathologie tumorale avec des malformations, et/ou syndromique de cancer héréditaire. De plus, ces pathologies sont caractérisées par un spectre tumoral spécifique et des caractéristiques phénotypiques qui leur sont propres et qui leur permettent d'être distinguées les unes des autres.

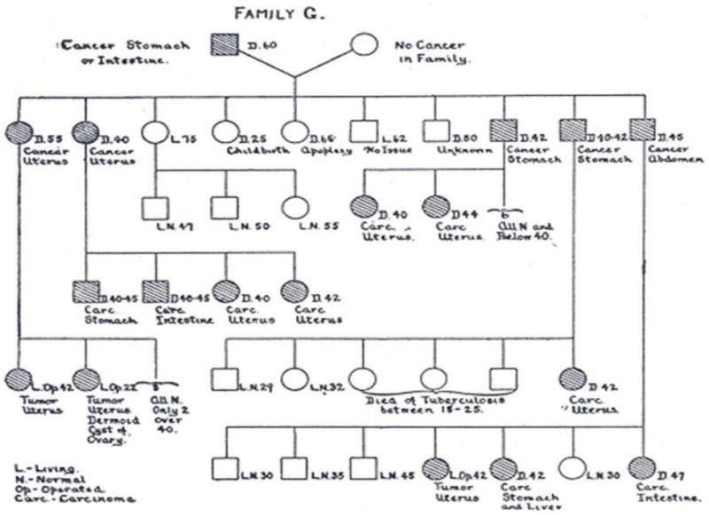
Parmi les syndromes de prédisposition héréditaire au cancer, le syndrome de Lynch ou cancer colorectal héréditaire non polyposique (HNPCC, *hereditary nonpolyposis colorectal cancer*) et le syndrome seins-ovaires (HBOC, *hereditary breast and ovarian cancer*) sont les plus fréquents avec une incidence de 1/400 et de 1/500-1/1000, respectivement.

2) La prédisposition génétique au cancer colorectal : le syndrome de Lynch

Le cancer colorectal (CCR) se définit comme une tumeur maligne qui prend naissance dans les cellules du côlon ou du rectum. Il s'agit de l'un des cancers les plus fréquemment diagnostiqué et représente la troisième forme de cancer la plus fréquente et la troisième cause de mortalité liée au cancer à travers le monde (pour revues : Favoriti *et al.*, 2016; Siegel *et al.*, 2017). Les CCR sporadiques, dépourvus de prédisposition familiale ou héréditaire, représentent 70% des CCR tandis que les 25% restant correspondent à des CCR avec une histoire familiale qui laisse suggérer une contribution génétique ou l'intervention des facteurs environnementaux ou une combinaison des deux (pour revue : Sameer, 2013). Ces formes familiales de CCR correspondent (i) majoritairement à des CCR familiaux, trop fréquents au sein d'une famille pour être considérés comme sporadiques mais dont la transmission n'est pourtant pas compatible avec un modèle héréditaire et (ii) à des prédispositions héréditaires au cancer du côlon, en particulier la polypose adénomateuse familiale (FAP, *familial adenomatous polyposis*) riches en polypes et le syndrome de Lynch (pour revue : Sameer, 2013).

Le syndrome de Lynch constitue la forme la plus fréquente de CCR héréditaires et représenterait environ 2 à 5 % de l'ensemble des CCR (pour revue : Sehgal *et al.*, 2014). Il s'agit également d'une des formes les plus anciennement décrites (pour revues : Lynch *et al.*, 2015; Sehgal *et al.*, 2014). En effet, il y a un siècle environ, le Dr Alfred Whartin, alors directeur du département de pathologie de l'université du Michigan a reporté, pour la première fois, cette maladie aujourd'hui dénommée syndrome de Lynch. En 1895, ce dernier est interpellé par le grand nombre de décès par cancer dans la famille de sa couturière, dans laquelle toutes les générations développent des cancers du côlon, de l'estomac ou de l'utérus à un âge jeune, y compris sa couturière qui succombera à un cancer de l'endomètre (Figure 2). Après avoir construit l'arbre généalogique de la famille G, ayant émigré depuis l'Allemagne (*Germany*) vers le Michigan, le Dr Warthin émit l'hypothèse que certains cancers pourraient être liés à un facteur héréditaire qui se transmettrait dans les familles. Depuis, plusieurs autres familles ont été rapportées, notamment les familles N (Nebraska) et M (Michigan) par Henri Lynch dans les années 70, confirmant ainsi l'hypothèse selon laquelle dans ces familles le "syndrome du cancer familial" (*cancer family syndrome*) était transmis selon une forme mendélienne et de manière autosomique dominante (pour revue : Lynch *et al.*, 2015). C'est en 1984 que Richard Boland attribuera le nom d'Henri Lynch au syndrome de Lynch (Boland and Troncale, 1984) et en 2000 que l'altération génétique causale retrouvée au sein de cette famille (MSH2 c.646-3T>G localisée sur le site accepteur d'épissage de l'exon 4 et entrainant une rétention de 24 nucléotides situés dans l'intron 3 en amont) a été mise en évidence (Yan *et al.*, 2000).

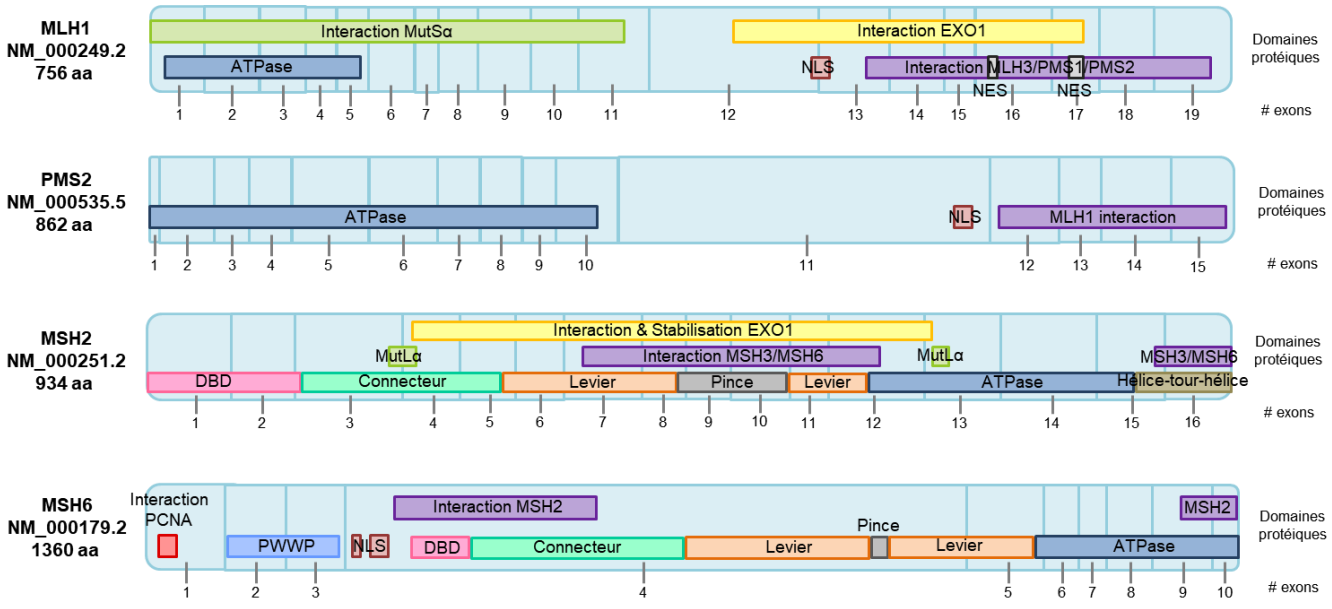
Figure 2 : Arbre généalogique de la Famille G (première famille évocatrice d'un syndrome de Lynch identifiée) initialement construit par le Dr Warthin (Warthin, 1895).



a. Les gènes MMR, gènes majeurs de prédisposition au syndrome de Lynch

Le syndrome de Lynch est une prédisposition génétique héréditaire qui obéit à un déterminisme mendélien selon un mode de transmission autosomique dominant, avec une pénétrance forte mais incomplète (environ 90%) et une expressivité variable. Quatre gènes majeurs de prédisposition ont été identifiés (Figure 3) : (i) *MLH1* (*MutL homolog 1 of E. coli* ; OMIM #120436), s'étendant sur 57.36 kb au niveau du chromosome 3 (3p21-23) et composé de 19 exons codant pour une protéine de 756 acides aminés (Lindblom *et al.*, 1993; Papadopoulos *et al.*, 1994; Peltomäki *et al.*, 1993), (ii) *MSH2* (*MutS homolog 2 of E. coli* ; OMIM #609309), s'étendant sur 80.1 kb au niveau du chromosome 2 (2p21) et composé de 16 exons codant pour une protéine de 934 acides aminés (Fishel *et al.*, 1993), (iii) *MSH6* (*MutS homolog 6 of E. coli* ; OMIM #600678) s'étendant sur 23.9 kb au niveau du chromosome 2 (2p16) et composé de 10 exons codant pour une protéine de 1361 acides aminés (Miyaki *et al.*, 1997), et (iv) *PMS2* (*postmeiotic segregation increased 2* ; OMIM #600259) s'étendant sur 16 kb au niveau du chromosome 7 (7p22) et composé de 15 exons codant pour une protéine de 862 acides aminés (Nicolaides *et al.*, 1994).

Figure 3 : Représentation des protéines MMR incluant les domaines fonctionnels et leurs correspondances avec les exons et les interacteurs protéiques (adapté de Bryony Thompson, communication personnelle). DBD, *DNA binding domain* ; PWWP, Pro-Trp-Trp-Pro ; NLS, *nuclear localization signal* ; NES, *nuclear export signal*.

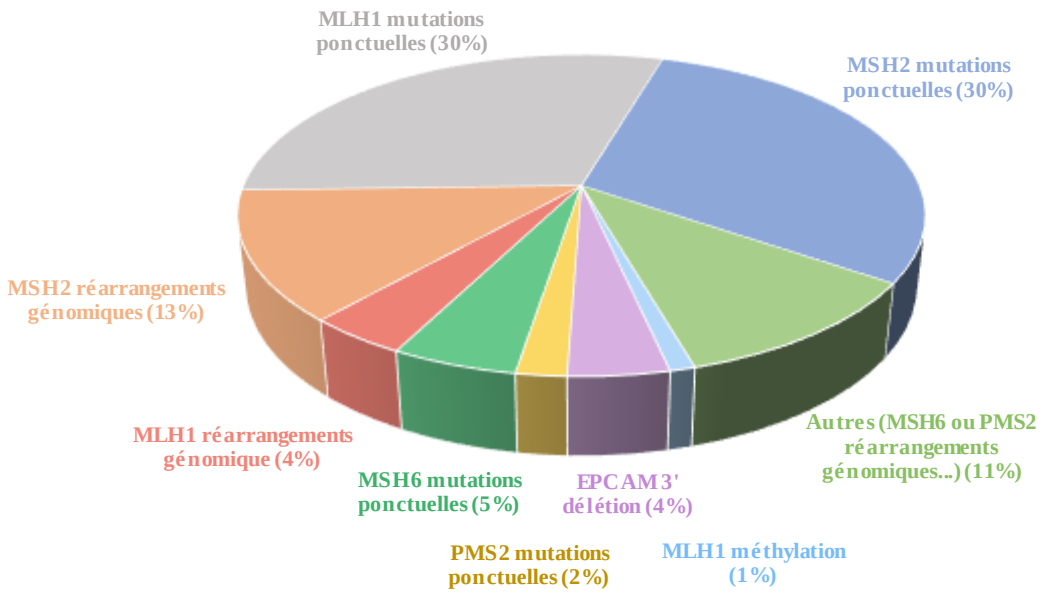


Ces gènes, dans leur ensemble appelés gènes MMR (*mismatch repair*), codent des protéines aux fonctions multiples, notamment connues pour leur rôle majeur dans le maintien de la stabilité du génome et de l'intégrité de l'information génétique au cours des multiples divisions cellulaire. La fonction essentielle du système MMR est la réparation post-réplicative des mésappariements de l'ADN, c'est-à-dire la réparation des erreurs de type mésappariements (présence dans la double hélice de deux bases non complémentaires) générés au moment de la réplication (mitose) par l'ADN polymérase (Section I.2.c) (pour revues : Jiricny, 2013; Sameer *et al.*, 2014). Outre la correction des mésappariements lors de la réplication, le système MMR a également une implication importante dans les phénomènes de recombinaison, échange d'une portion d'ADN entre deux brins provenant de différents duplexes, survenant en particulier lors de la méiose (pour revue : Spies and Fishel, 2015). Lorsque les deux séquences d'ADN ne sont pas identiques (recombinaison hétérologue), il y a formation de mésappariements. Ces derniers peuvent être reconnus puis corrigés par le système MMR qui va alors remplacer l'information contenue sur un chromosome avec celle présente sur le chromosome avec lequel il recombine. On parle alors de conversion génique. Si la réparation du mésappariement est impossible, le système MMR peut faire avorter le processus de recombinaison. On parle alors d'anti-recombinaison. Il a été suggéré que ce phénomène pourrait faire intervenir des hétérodimères MSH (MSH4/MSH5), qui en interagissant avec les protéines impliquées dans l'échange de brins (RecA), bloquent leur fonction et inhibent ainsi la recombinaison (pour revue : Spies and Fishel, 2015). Ce mécanisme d'anti-recombinaison est d'une grande importance sur le plan du maintien de l'intégrité du génome et joue un rôle déterminant dans l'évolution. De plus, les protéines du système MMR interviennent dans le contrôle du cycle cellulaire, notamment au niveau du point de contrôle G2/M, pendant lequel elles peuvent induire l'arrêt du cycle cellulaire et éventuellement l'apoptose suite à des altérations de l'ADN autres que les mésappariements, induites par des agents chimiques comme les drogues antinéoplasiques, la cisplatine et la carboplatine et les agents alkylants tels que la 6-thioguanine (O'Brien and Brown, 2006). En effet, il a été montré que les cellules déficientes en protéines MMR étaient 100 fois plus résistantes aux agents alkylants et 2 à 4 fois plus résistantes à la cisplatine (O'Brien and Brown, 2006). Les protéines MMR sont également impliquées dans la réparation des lésions oxydatives de type 8-oxoguanine (pour revue : Martin *et al.*, 2010).

b. Les bases moléculaires du syndrome de Lynch

Dans 70% des cas, le syndrome de Lynch résulte d'une altération germinale hétérozygote d'un des gènes du système MMR, dont l'implication respective n'est pas équivalente (Olschwang and Eisinger, 2010). En effet, les risques cumulés de développer une tumeur du spectre du syndrome de Lynch en fonction de l'âge varie en fonction du gène en cause tel qu'à 70 ans cette incidence est estimée à 57, 59 et 25% lorsque *MSH2*, *MLH1* ou *MSH6* sont impliqués, respectivement (Bonadona *et al.*, 2011). Une contribution majeure est attribuée aux gènes *MLH1* et *MSH2*, qui expliquent à eux seuls plus de 75% des cas avec une mutation pathogène identifiée (Figure 4 ; Grandval *et al.*, 2013; Plazzer *et al.*, 2013).

Figure 4 : Contribution respective des gènes MMR au syndrome de Lynch (d'après Olschwang and Eisinger, 2010). L'implication des différentes altérations constitutionnelles des gènes MMR a été mise à jour en 2009 dans une analyse synthétique de l'ensemble des études publiées (Sheng *et al.*, 2009).

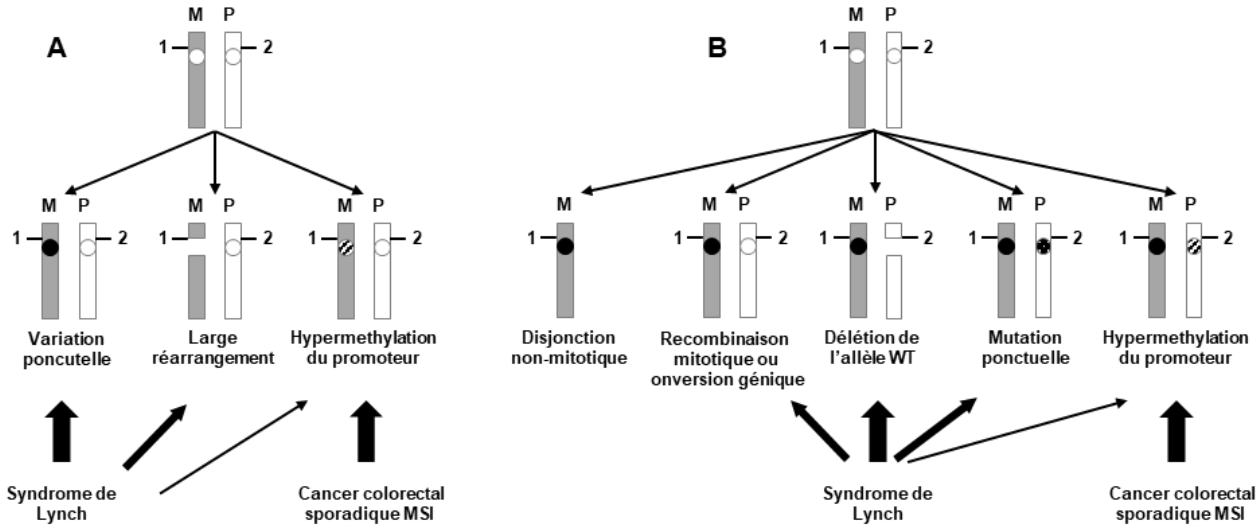


Alors que des mutations ponctuelles hétérozygotes sont trouvées en proportion égale sur les deux gènes *MSH2* et *MLH1*, à savoir 30 à 35 % des mutations pathogènes associées au syndrome de Lynch, les réarrangements génomiques de type délétions ou duplications sont trouvés dans environ 17 % des cas, quatre fois plus fréquemment sur *MSH2* (13%) que sur *MLH1* (4%) (pour revue : Olschwang and Eisinger, 2010). Dans une moindre mesure, des mutations pathogènes,

beaucoup moins fréquentes, sont associées aux gènes *MSH6* et *PMS2* et représentent 18% des mutations pathogènes reportées (pour revue : Olschwang and Eisinger, 2010). Plus rarement (1% des cas), certaines tumeurs associées au syndrome de Lynch sont caractérisées par la présence d'une modification épigénétique constitutionnelle dans *MLH1* (pour revues : Olschwang and Eisinger, 2010; Sehgal *et al.*, 2014). Il s'agit plus exactement d'une hyperméthylation du promoteur du gène qui conduit l'inhibition de la transcription de l'allèle concerné et qui est considérée comme un mécanisme alternatif à l'origine du syndrome de Lynch. Récemment, des délétions constitutionnelles spécifiques dans la partie 3' du gène *EPCAM* (4% des mutations), situé en amont du gène *MSH2* ont été décrites comme à l'origine du syndrome de Lynch par inactivation épigénétique du gène *MSH2* (pour revue : Olschwang and Eisinger, 2010; Sehgal *et al.*, 2014). En effet, la délétion de la partie 3' contenant le codon stop du gène *EPCAM* entraine une « inactivation somatique héritable de *MSH2* ». A noter que ces données varient en fonction des bases de données. En effet, au niveau national, il a été rapporté que 36% des mutations associées au syndrome de Lynch sont localisées dans *MLH1*, 39% dans *MSH2* et 25% dans *MSH6* (Grandval *et al.*, 2013), tandis qu'au niveau international 42% des variations détectées dans le syndrome de Lynch sont localisées dans *MLH1*, 33% dans *MSH2*, 18% dans *MSH6* et 7% dans PMS2 (Plazzer *et al.*, 2013).

Les gènes MMR codent pour des protéines intervenant dans une des voies de réparation de l'ADN assurant ainsi le maintien de la stabilité et de l'intégrité du génome. Par conséquent, ces gènes sont considérés comme des gènes de stabilité du génome (*caretakers*), et plus largement, comme des gènes supresseurs de tumeur. Les gènes MMR répondent donc au modèle à deux coups (*2 hits*) de Knudson dans lequel une perte de fonction des protéines codées par ces gènes nécessite l'inactivation des deux allèles du gène pour conduire à l'apparition d'une tumeur par accumulation d'altérations génétiques non réparées (pour revues : Friedberg, 2003; Knudson, 1971). Le premier évènement correspond à une altération constitutionnelle, présente dès la naissance, dans l'un des gènes MMR, altération de type variation ponctuelle le plus souvent ou de type grand réarrangement. Le second événement correspond quant à lui à une altération somatique qui parvient dans les cellules progénitrices du tissu cancéreux (pour revue : Peltomäki, 2014). Il peut s'agir d'une variation ponctuelle, d'une délétion ou d'une hyperméthylation du promoteur, ou encore d'une conversion génique (Figure 5).

<u>Figure 5 :</u> Comparaison des mécanismes moléculaires à l'origine des *2 hits* dans les cancers associés au syndrome de Lynch et les cancers sporadiques avec instabilité microsatellitaire (d'après Peltomäki, 2014)**.** (A) Mécanismes moléculaires correspondant à la première mutation (*first hit*). (B) Mécanismes moléculaires correspondant à la seconde mutation (*second hit*) P, paternel ; M, maternel ; 1 et 2 correspondent au 1er et 2ème allèle de *MLH1* respectivement. L'épaisseur de la flèche indique la signification relative de chaque mécanisme. L'allèle maternel a été choisi d'une manière arbitraire pour être porteur de la première mutation.



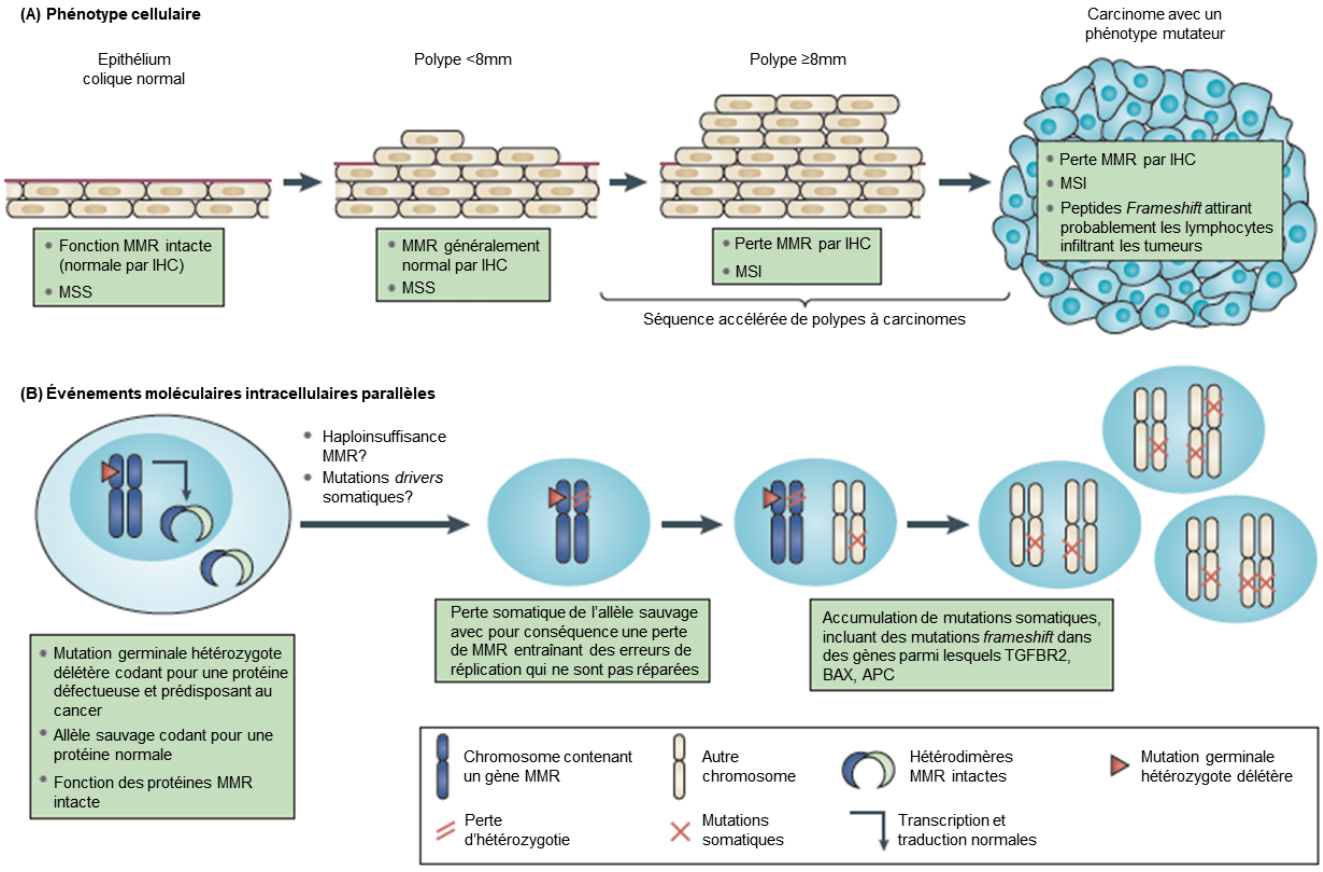
Suite à l'inactivation des 2 allèles, la voie de réparation impliquant le système MMR devient complètement inactive. Or, le système MMR est impliqué dans la réparation des erreurs de type mésappariements de l'ADN survenant lors de la réplication (pour revues : Jiricny, 2013; Sameer *et al.*, 2014). De ce fait, la déficience du système MMR induit une accumulation des erreurs lors de la réplication de l'ADN qui se traduit par une augmentation importante du taux de mutations dans les cellules déficientes comparativement aux cellules normales. Ceci a pour conséquence d'augmenter drastiquement l'instabilité génomique, particulièrement au niveau des microsatellites, séquences d'ADN mono, di, tri ou tétra-nucléotidiques répétées en tandem et distribuées au sein du génome. En effet, l'inactivation du système MMR conduit à l'accumulation d'insertions ou de délétions (indels) dans la séquence répétée des microsatellites, principalement retrouvées au niveau de l'ADN issu des cellules tumorales comparativement à l'ADN issus des cellules du tissu normal et qui ne peuvent être réparées (pour revues : Jiricny, 2013; Sameer *et al.*, 2014). On parle alors d'instabilité microsatellitaire (MSI, *<u>micro</u><u>s</u><u>atellite</u> <u>i</u>nstability*) qui, *in fine*, contribue à l'apparition

d'altérations secondaires susceptibles d'activer les proto-oncogènes, d'inactiver les *gatekeepers* voire même d'autres *caretakers* (pour revue : Vogelstein and Kinzler, 2015). En particulier, l'instabilité somatique des séquences répétées entraîne l'inactivation somatique de gènes impliqués dans la carcinogenèse colorectale par décalage du cadre de lecture. On parle cette fois-ci de phénotype RER (*replication error*). Parmi ces gènes, on trouve le gène codant pour le récepteur de type II du TGFβ (*TGFβRII*), le gène *APC*, le gène du récepteur de l'IGFII (*IGFIIR*), les deux gènes MMR humains *hMSH3* et *hMSH6* homologues de MutS, et le gène de la protéine pro-apoptotique BAX (pour revues : Lynch *et al.*, 2015; Peltomäki, 2001). Cette séquence d'inactivation explique pourquoi les formes héréditaires de cancer résultant de mutations constitutionnelles inactivatrices de *caretakers* comme les gènes MMR se caractérisent généralement par une pénétrance incomplète (probabilité d'être atteint par la maladie, à un moment donné) et une survenue plus précoce des tumeurs comparativement aux tumeurs sporadiques. Une déficience du système MMR a également été observée dans d'autres cancers héréditaires alors considérés comme des variantes rares du syndrome de Lynch (pour revue : Buecher and Laurent-Puig, 2010). Il s'agit notamment des syndromes de Muir-Torre, caractérisé par l'association de lésions sébacées avec des cancers de type CCR le plus souvent (Muir *et al.*, 1967; Torre, 1968) et de Turcot, caractérisé par une association de tumeurs cérébrales et de CCR dans une même famille (Turcot *et al.*, 1959). De plus, des variations constitutionnelles à l'état homozygote sont associées au CMMR-D (*constitutional MMR deficiency syndrome*) (Ricciardone *et al.*, 1999; Wang *et al.*, 1999) ainsi que dans des cancers sporadiques (Hsieh and Yamane, 2008).

Cette séquence d'événements moléculaires intracellulaires que subissent les cellules en transformation conduit à la progression du phénotype cellulaire nécessaire au développement du CCR (Figure 6 ; pour revue : Lynch *et al.*, 2015). En effet, bien que porteurs d'une mutation constitutionnelle hétérozygote dans un gène MMR, le tissu épithélial du côlon reste néanmoins normal, avec un statut MSS (*microsatellite stable*) et une expression protéique normale des quatre protéines en immunohistochimie (IHC). Survient alors le développement de polypes. Le processus néoplasique qui conduit à la formation de polypes n'est pour l'instant pas connu mais il pourrait faire intervenir des mutations somatiques *drivers*. Ce n'est que lorsque le diamètre de ces polypes dépasse 8 mm que le statut MSI et la perte de protéine MMR sont détectés, consécutivement à la perte de fonction de l'allèle sauvage au niveau somatique. L'inactivation du système MMR entraîne l'accumulation de mutations somatiques dans plusieurs gènes, notamment les gènes suppresseurs

de tumeur *TGFBR2*, *BAX* et *APC*, qui participent à la progression rapide de la tumeur. De plus, l'instabilité somatique des séquences répétées conduit à un décalage du cadre de lecture au sein de ces gènes, qui génère des peptides dits *frameshift* ou neopeptides. Certains de ces peptides sont exprimés à la surface de la cellule et vont alors être reconnus comme antigènes spécifiques de la tumeur par un grand nombre de lymphocytes infiltrant la tumeur qui vont ainsi induire la réaction immunitaire.

Figure 6 : Développement du cancer colorectal dans le cadre d'un syndrome de Lynch (d'après Lynch *et al.*, 2015). A) Evolution du phénotype cellulaire de l'épithélium colique en présence d'une mutation constitutionnelle dans l'un des gènes *MMR*. B) Evènements moléculaires intracellulaires parallèles. IHC, Immunohistochimie ; MMR, *mismatch repair* ; MSI, *microsatellite instability* ; MSS, *microsatellite stable*.

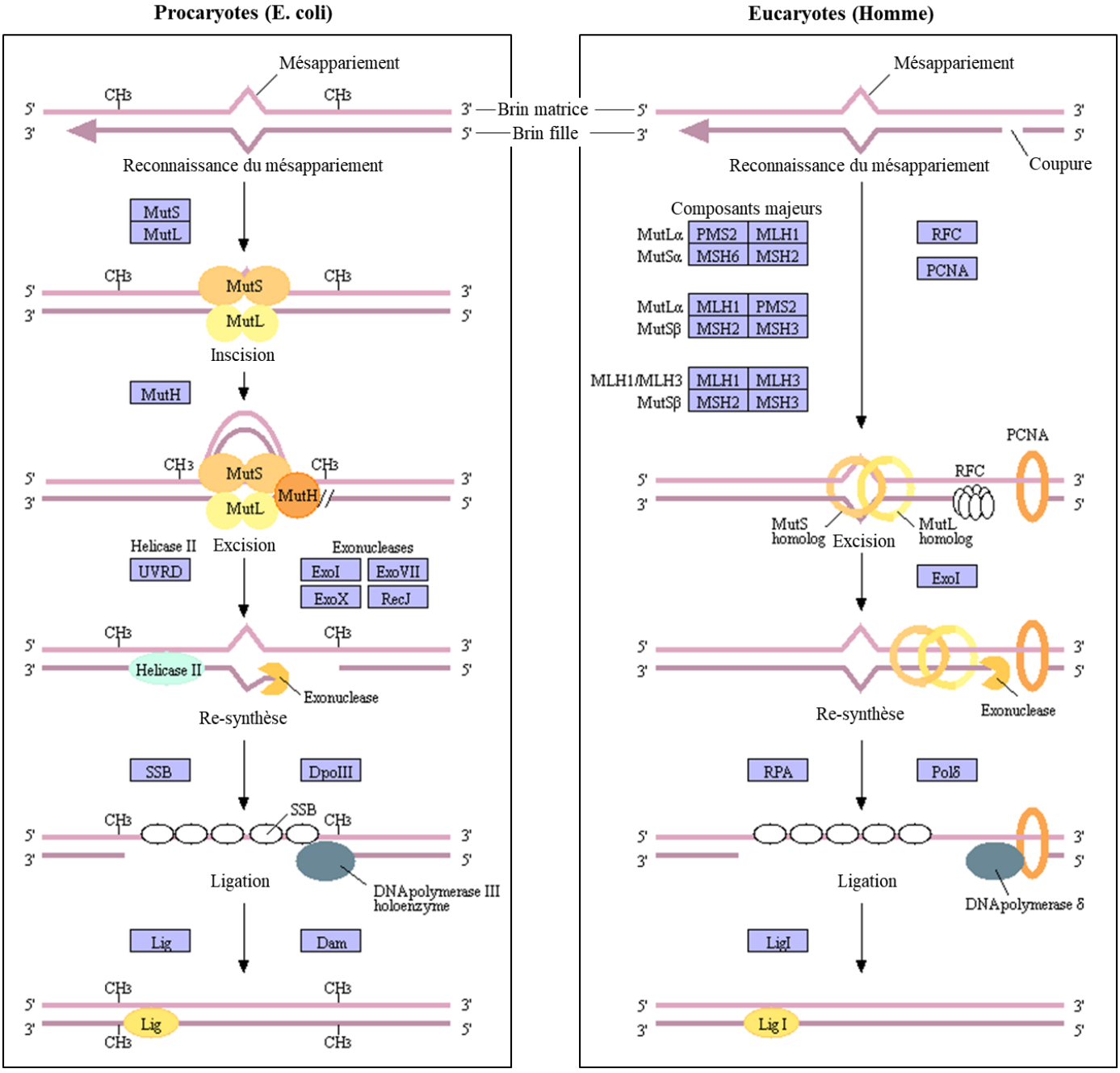


c. Le système de réparation des mésappariements de l'ADN

L'acide désoxyribonucléique (ADN) est la molécule à la base du vivant : elle contient l'information nécessaire au développement et au fonctionnement de la cellule, tout en garantissant la transmission de cette information aux générations suivantes, constituant ainsi l'hérédité. Le maintien de l'intégrité de cette information génétique est essentiel et est assuré principalement par l'efficacité du processus de réplication de l'ADN, grâce à la fidélité des ADN polymérases. Ces enzymes incorporent les bases nucléiques en fonction de la complémentarité des appariements Watson-Crick (A-T et C-G), avec un taux d'erreur relativement faible, d'environ 10^{-5} (une erreur sur cent mille bases répliquées). De plus, les ADN polymérases disposent d'une activité de relecture (*proofreading*) ou activité exonucléasique 3'→5', qui leur permet de corriger leurs propres erreurs lors de la réplication, en éliminant le dernier nucléotide incorporé pour ensuite reprendre la synthèse du brin d'ADN. L'étape de relecture fait chuter le taux d'erreur de réplication à environ 10^{-7}. Pourtant, des erreurs de mésappariements (*mismatch repair*) peuvent éventuellement échapper au mécanisme de contrôle par les polymérases. C'est pourquoi la cellule dispose d'un mécanisme spécifique de surveillance et de réparation des lésions réplicatives, notamment les mésappariements dans l'ADN double brin, indépendant des polymérases. Grace à ce système, le taux d'erreur de réplication chute à environ une mutation pour 10 milliards de bases répliquées et contribue avec les ADN polymérases au maintien de l'intégrité de l'information génétique (pour revue : Kunkel and Erie, 2015). Le système MMR fait intervenir un ensemble de protéines intervenant dans la détection, l'excision et la resynthèse d'une partie du brin néosynthétisé contenant des erreurs de type mésappariements de bases ou de type boucles insertions/délétions de bases (Figure 7).

Le système MMR le mieux caractérisé est celui de la bactérie *Escherichia coli* (E. *coli*). Ce système appelé système MutHLS (Mut, m*utator*) agit selon un processus post-réplicatif d'excision-resynthèse d'une portion d'ADN contenant un mésappariement (Figure 7 ; pour revues : Iyer *et al.*, 2006; Reyes *et al.*, 2015). Il se compose de trois protéines principales : MutS, MutL et MutH, formant le système MutHLS. La réparation des mésappariements de l'ADN par le système MMR débute tout d'abord avec la reconnaissance du mésappariement par un homodimère de protéines MutS. Cette reconnaissance du mésappariement permet le recrutement de l'homodimère de protéines MutL, initiant ainsi la réaction de réparation. Ensuite, l'endonucléase MutH est recrutée à son tour par MutL puis activée. Cette dernière est capable de reconnaitre puis de cliver

<u>Figure 7 :</u> Description et comparaison du système de réparation des mésappariements de l'ADN chez les procaryotes (*Escherichia coli)* et les eucaryotes (Homme) (d'après Kanehisa Laboratories).



spécifiquement le brin d'ADN néosynthétisé grâce à son activité endonucléase. En effet, cette enzyme est une endonucléase sensible à la méthylation, modification post-réplicative assurée par l'enzyme DAM (désoxyadénosine méthylase) et retrouvée au niveau de l'ADN bactérien des gammaproteobacteria auquel appartient E. *coli*. Plus précisément, le brin d'ADN néosynthétisé est facilement détectable puisqu'il est non méthylé de manière transitoire, contrairement au brin parental. Ainsi le complexe MutH génère de façon exclusive une coupure au niveau d'une séquence GATC non méthylée, située en 3' ou en 5' du mésappariement dans le nouveau brin d'ADN. Il s'ensuit alors une réaction d'excision, survenant au niveau de la coupure générée dans le brin néosynthétisé. Le processus d'excision requiert la présence de l'ADN hélicase II (MutU ou UvrD) qui va dérouler l'ADN et d'exonucléases, ExoVII et RecJ ou ExoI, ExoVII et ExoX, qui vont quant à elle digérer, de 5' en 3' ou de 3' en 5', respectivement, le nouveau brin d'ADN contenant le mésappariement. Ce brin va alors être stabilisé dans un premier temps par les protéines de liaison à l'ADN simple brin ou SSB (*single strand DNA binding protein*) puis re-synthétisé par l'holoenzyme ADN polymérase III. L'ADN sera finalement recircularisé grâce à l'ADN ligase, qui va rétablir la continuité covalente du brin réparé après l'élongation effectuée par l'ADN polymérase (pour revues : Iyer *et al.*, 2006; Reyes *et al.*, 2015). De manière intéressante, l'inactivation du système MutHLS entraîne un taux accru (x1000) de mutations spontanées (Glickman and Radman, 1980).

Chez les eucaryotes, le mécanisme de réparation des mésappariements de l'ADN est similaire à celui des bactéries. En effet, le système MMR est très conservé de la bactérie aux mammifères. Ainsi, il existe de très fortes homologies entre les systèmes MMR de la bactérie, de la levure et de l'homme (pour revue : Reyes *et al.*, 2015). Chez l'homme, les principaux complexes MMR sont essentiellement composés de protéines homologues aux protéines MutS et MutL bactériennes, appelés respectivement MSH (*MutS homolog*) et MLH (*MutH homolog*). Mais contrairement à leurs homologues MutS et MutL présents sous la forme d'homodimères chez E. *coli*, les complexes MSH et MLH sont retrouvés sous la forme d'hétérodimères chez les eucaryotes, grâce à des domaines d'interaction spécifiquement contenus dans ces protéines (Figure 7 ; Iyer *et al.*, 2006; Reyes *et al.*, 2015). En effet, l'hétérodimère MutSα est constitué des protéines MSH2 et MSH6 (complexe MSH2-MSH6) reconnaissant essentiellement, via leur domaine de liaison à l'ADN, les mésappariements d'ADN de type substitutions ponctuelles tandis que l'hétérodimère MutSß se compose des protéines MSH2 et MSH3 (complexe MSH2-MSH3) spécialisées dans la

reconnaissance des mésappariements d'ADN de type indels. A noter que chez l'homme, le complexe MSH2-MSH6 est 10 fois plus abondant que MSH2-MSH3. Les hétérodimères MLH se composent quant à eux des protéines MLH1 et PMS2 formant le complexe MuLα, le plus connu chez l'homme, et des protéines MLH1-PMS1 et MLH1-MLH3, formant, respectivement, les complexes additionnels MutLß et MutLγ, ayant un rôle mineur dans le système MMR des eucaryotes. Contrairement à E. *coli*, la protéine MutH est absente chez les eucaryotes. En effet, le domaine C-terminal de certains complexes MLH, en particulier MutLα et MutLγ, contiennent un motif endonucléase bien conservé qui n'est pas présent chez E. *coli*. Ainsi, ces complexes MLH sont dotés d'une activité endonucléase intrinsèque, qui explique l'absence du complexe MutH chez les eucaryotes. En plus des hétérodimères MLH et MSH, des protéines RPA (*replication protein A*), protéine de liaison à l'ADN simple brin qui stabilise le brin d'ADN non-endommagé sous forme monocaténaire, de la ligase, de l'exonucléase 1 et des polymérases δ et ε retrouvées sous la forme d'homologue chez E. *coli*, il a été montré que les facteurs RFC (*replication factor C*) et PCNA (*proliferating cell nuclear antigen*), interviennent également dans le processus d'excision bidirectionnelle chez les eucaryotes. Mais contrairement à ce qui a pu être décrit chez les procaryotes, le choix du brin à réparer ne repose pas sur la reconnaissance des sites méthylés. D'autres mécanismes de discrimination du brin porteur du mésappariement ont été évoqués. Il a été suggéré que, chez les eucaryotes, l'ADN nouvellement synthétisé contiendrait, de façon transitoire, des coupures, avant avant d'être scellées par la ligase. Ces coupures fourniraient un signal qui dirige le système MMR vers le brin approprié. PCNA pourrait jouer un rôle important dans ce contexte (pour revues : Iyer *et al.*, 2006; Reyes *et al.*, 2015).

d. Critères moléculaires et cliniques évocateurs d'un syndrome de Lynch

Le diagnostic du syndrome de Lynch repose sur un diagnostic moléculaire, axé sur la recherche, dans le génome d'un patient, d'une mutation constitutionnelle hétérozygote délétère ou d'une épimutation entrainant la perte de fonction de l'un des gènes MMR (pour revue : Lynch *et al.*, 2015). Ce diagnostic moléculaire consiste d'abord en la recherche de mutations ponctuelles sur les régions codantes et les régions introniques flanquantes des gènes *MSH2* et *MLH1* par séquençage haut-débit. Si aucune mutation ponctuelle n'est détectée, la recherche de réarrangements génomiques dans *MSH2*, *MLH1* ainsi *qu'EPCAM* est effectuée par la technique de

MLPA (*multiplex ligation-dependent probe amplification*). Enfin, si aucune mutation ponctuelle ou aucun réarrangement ne sont détectés dans les gènes *MSH2* et *MLH1*, les gènes *PMS2* et *MSH6* sont alors analysés. Concernant la recherche de mutation dans le gène *PMS2*, il est important de noter qu'il existe plusieurs pseudogènes de *PMS2*, y compris sur le même chromosome, ce qui rend difficile la détection des variations dans ce gène (Nakagawa *et al.*, 2004). C'est le cas notamment du pseudogène *PMS2CL*, situé à proximité de *PMS2* (7p), qui partage 98% d'homologie avec *PMS2* au niveau des exons 9 et 11-15. Une méthode de PCR long-range, spécifique de *PMS2* uniquement a donc été développée (Vaughn *et al.*, 2010).

Le diagnostic moléculaire n'est réalisé qu'après identification d'un certain nombre de signes évocateurs de cette maladie, en particulier les caractéristiques tumorales et les antécédents personnels et/ou familiaux, qui vont orienter la recherche de la variation constitutionnelle dans l'un des gènes MMR (Tableau 2). L'une des caractéristiques déterminant le diagnostic moléculaire est le spectre tumoral observé chez les individus porteurs. En effet, le syndrome de Lynch est une prédisposition génétique héréditaire touchant principalement le côlon et l'endomètre. Outre les CCR et de l'endomètre, le spectre tumoral du syndrome de Lynch peut également être associé aux cancers de l'intestin grêle, des voies urinaires excrétrices (uretère et bassinet) formant ce qui est appelé le spectre tumoral étroit du syndrome de Lynch (pour revue : Lynch *et al.*, 2015). Le spectre tumoral peut également être élargi aux cancers de l'estomac, des ovaires, du pancréas, des voies biliaires, du système nerveux central (glioblastome : syndrome de Turcot), et adénomes des glandes sébacées et kératoacanthomes (syndrome de Muir-Torre) (pour revue : Lynch *et al.*, 2015). Certaines de ces tumeurs sont préférentiellement associées à des altérations dans un gène donné, donnant lieu à une hétérogénéité phénotypique (Tableau 3). En effet, il a été montré que les individus porteurs d'une mutation dans le gène *MSH2* développent de manière prépondérante des cancers extracoliques tandis qu'ils développent moins fréquemment des CCR comparativement aux individus porteurs d'une mutation dans le gène *MLH1* (Lin *et al.*, 1998; Vasen *et al.*, 2001). De même, les mutations du gène *MSH6* sont très communément liées aux cancers gastro-intestinaux et de l'endomètre avec un âge de présentation plus avancé (Buchanan *et al.*, 2014; Hendriks *et al.*, 2004). D'ailleurs, les mutations du gène *MSH6* sont considérées comme une cause fréquente du syndrome de Lynch dit atypique, car il ne remplit pas totalement les critères d'identification d'Amsterdam (Tableau 2; Buchanan *et al.*, 2014; Hendriks *et al.*, 2004).

Tableau 2 : Description des critères d'Amsterdam I et II et des critères de Bethesda (adapté de Lynch *et al.*, 2015). CRC, cancer colorectal ; FAP, polypose adénomateuse familiale ; LS, Syndrome de Lynch

Critères d'Amsterdam I et II (Très sélectifs)	**Critères de Bethesda (Moins sélectifs)**
Au moins trois apparentés avec un CRC (Amsterdam I) ou un cancer du spectre étroit du syndrome de Lynch (Amsterdam II) histologiquement prouvé **et :** 1) Un des patients est apparenté au premier degré avec les deux autres 2) Au moins deux générations successives sont atteintes 3) Au moins un des cas a été diagnostiqué avant l'âge de 50 ans 4) La FAP doit être exclue dans chacun des cas de CRC	1) CCR diagnostiqué avant l'âge de 50 ans **ou** 2) CCR synchrone, métachrone ou associé à un autre cancer appartenant au spectre du SL quel que soit l'âge **ou** 3) CCR avec des caractéristiques anatomo-pathologiques évocatrice (infiltrat lymphocytaire dense du stroma tumoral, réaction inflammatoire de type Crohn, différenciation mucineuse ou en bague à chaton, architecture de type médullaire) diagnostiqué avant l'âge de 60 ans **ou** 4) CCR avec au moins un apparenté au premier degré atteint d'un cancer appartenant au spectre du syndrome de Lynch*, une des tumeurs est diagnostiquée avant l'âge de 50 ans **ou** 5) CCR avec au moins deux apparentés au premier ou au second degré avec un cancer appartenant au spectre du syndrome de Lynch, quel que soit l'âge.

Tableau 3 : Hétérogénéité phénotypique associée au syndrome de Lynch en fonction de la variation constitutionnelle détectée (adapté de Lynch *et al.*, 2015). CCR, cancer colorectal ; FAP, polypose adénomateuse familiale ; MSI, instabilité microsatellitaire.

Mutations	**Hétérogénéité phénotypique associée**
Mutation hétérozygote *MLH1*	Prédominance des CRC Cancers extracoliques moins fréquents qu'avec des mutations *MSH2*
Mutation hétérozygote *MSH2*	Fréquence élevée de cancers extracoliques
Mutation hétérozygote *MSH6*	Prédominance des cancers endométriaux Certaines tumeurs présentent une MSI faible
Mutation hétérozygote *PMS2*	Peut être associée à un excès de polypes coliques Cancer moins fréquent
Délétion hétérozygote *EPCAM*	Perte d'expression de *MSH2*. Souvent un plus faible risque de cancers extracoliques, mais si la délétion est proche du gène *MSH2*, le risque de cancer endométrial augmente
Epimutation monoallélique *MLH1*	Expression phénotypique similaire aux porteurs de mutation *MLH1* Majoritairement sporadique (*de novo*)

Outre le spectre tumoral, une signature moléculaire spécifique des tumeurs associées au Syndrome de Lynch peut être mise en évidence chez les cas évocateurs d'un syndrome de Lynch ou de survenue d'un cancer colorectal ou de l'endomètre à un âge précoce, en particulier l'instabilité microsatellitaire. Celle-ci peut être mise en évidence à l'aide d'un criblage par PCR (pour revue : Buecher *et al.*, 2011). Il s'agit d'amplifier un certain nombre de microsatellites à la recherche d'une instabilité génétique, témoignant le manque de correction de d'erreurs de réplication de l'ADN au niveau tumoral. Cette technique est basée sur l'amplification de 5 « marqueurs » microsatellites. Alors que les microsatellites étudiés, il y a quelques années correspondaient à 2 marqueurs mono-nucléotidiques (BAT-25 et BAT-26) et à 3 marqueurs di-nucléotidiques (D2S123, D5S346, D17S250), selon les recommandations de l'Institut National de Cancer américain (Boland *et al.*, 1998; Richman, 2015), l'étude porte actuellement sur 5 marqueurs mono-nucléotidiques (BAT-25 localisé dans l'intron 16 du gène *C-KIT* ; BAT-26 localisé dans l'intron 5 du gène *MSH2* ; NR-21 localisé dans la région 5' non traduite du gène *SLC7AB* ; NR-24 localisé dans la région 5' non traduite du gène *ZNF-2* et NR-27 (MONO-27) localisé dans la région 5' non traduite du gène *IAP-1*) dont la spécificité est supérieure à celle des répétitions de di-nucléotides pour le diagnostic d'instabilité des microsatellites. De plus, ces microsatellites sont quasi-monomorphes, c'est-à-dire qu'ils sont caractérisés par un nombre de répétitions, et donc une taille, très homogène dans une population donnée (entre les différents individus et pour les 2 allèles d'un même individu). Les polymorphismes (SNP, *single nucleotide polymorphism*) ne sont cependant pas exclus, en particulier pour les marqueurs BAT-25 et NR-21. Trois phénotypes peuvent être observés après criblage : (i) instabilité microsatellitaire élevée (MSI-H, MSI-high) ou plus rarement dMMR (*deficient MMR*) pour les cellules tumorales présentant plus de 20% de variations microsatellitaires (c'est-à-dire positives pour au moins deux marqueurs sur les 5 microsatellites analysés) ; instabilité microsatellitaire faible (MSI-L, *MSI-low*) pour les cellules tumorales présentant une instabilité microsatellitaire inférieure à 20% (c'est-à-dire positives pour un marqueur sur les 5 marqueurs analysés) et microsatellites stables (MSS) ou pMMR (*proficient MMR*) lorsque qu'aucune variation microsatellitaire n'est détectée (c'est-à-dire négatives pour l'ensemble des marqueurs microsatellitaires) (pour revue : Sehgal *et al.*, 2014). Toutefois, même si la quasi-totalité des cancers survenant dans le contexte du syndrome de Lynch sont associés à une instabilité microsatellitaire, il est estimé qu'environ 15% des CCR sporadiques présentent également un phénotype MSI. Le mécanisme en cause est une hyperméthylation acquise du

promoteur du gène *MLH1*, engendrant une perte d'expression de la protéine MLH1 (éventuellement associée à un défaut d'expression de sa protéine partenaire PMS2). Dans ce cas, un génotypage de la variation somatique *BRAF* V600E combiné à l'occurrence au niveau somatique d'un phénotype MSI peut confirmer le caractère sporadique du CCR et exclure le diagnostique d'un syndrome de Lynch (Buecher *et al.*, 2011).

De plus, les tumeurs associées au syndrome de Lynch sont également associées à la perte d'expression totale d'au moins une des protéines MMR (MLH1, MSH2, MSH6, PMS2). Afin d'analyser l'expression protéique des 4 composants principaux du système MMR dans les tissus tumoraux, des études immunohistochimiques, méthode rapide et simple, sont généralement réalisées (pour revue : Richman, 2015). Dans la plupart des cas, les tumeurs déficientes en activité MMR montrent une extinction totale d'au moins une des protéines MMR étant donné que la protéine normalement codée par le gène muté n'est généralement pas exprimée (pour revue : Richman, 2015). Ainsi, le résultat de l'étude immunohistochimique donne une indication sur le gène en cause et permet d'orienter l'étude moléculaire constitutionnelle. Par exemple, une perte d'expression isolée de PMS2 ou MSH6 au niveau tumoral suggère la présence d'une mutation constitutionnelle dans les gènes *PMS2* et *MSH6*, respectivement. Cependant, concernant les protéines MLH1 et MSH2, une perte d'expression conjointe avec les protéines PSM2 et MSH6 est souvent observée. En effet, lors de la réaction de réparation des mésappariements de l'ADN par le système MMR, les protéines MMR fonctionnent sous forme d'hétérodimères MLH1/PMS2 et MSH2/MSH6. Ainsi, un défaut d'expression simultané des protéines MSH2 et MSH6 ou MLH1 et PMS2, suggère la présence potentielle d'une mutation constitutionnelle de *MSH2* et *MLH1*, respectivement. Cela s'explique par la déstabilisation des protéines MSH6 et PMS2 en l'absence des protéines MSH2 et MLH1, respectivement (pour revues : Lynch *et al.*, 2015; Richman, 2015).

Des recommandations internationales de bonnes pratiques (*guidelines*) standardisées, telles que les critères d'Amsterdam I et II, de Bethesda et de Bethesda modifiés, ont d'ailleurs été établies à partir de ces signes évocateurs cliniques et moléculaires pour faciliter le diagnostic du syndrome de Lynch (Tableau 2). Les critères d'identification dits d'Amsterdam I, définis en 1991 par le groupe collaboratif international du syndrome de Lynch, permettent d'identifier les individus, avec une histoire familiale de CCR, présentant un risque accru de développer un CCR à un âge précoce, tout en excluant le syndrome de FAP (pour revues : Lynch *et al.*, 2015; Richman, 2015). En 1999,

les critères d'Amsterdam II, version modifiée des critères d'Amsterdam I, ont été établis afin d'inclure le concept de tumeurs extracoliques, élargissant ces critères à d'autres localisations tumorales que le colon (Vasen *et al.*, 1999). Ces critères, bien que très spécifiques, restent assez peu sensibles (pour revue : Vasen, 2007) d'où la nécessité d'une révision qui a conduit aux critères de Bethesda (Rodriguez-Bigas *et al.*, 1997). A la différence des critères d'Amsterdam, les critères de Bethesda, établis en 1996 par l'Institut National du Cancer américain, incluent l'instabilité microsatellitaire comme une signature des tumeurs associées au syndrome de Lynch. Comme pour les critères d'Amsterdam I, les critères de Bethesda ont été réévalués puis modifiés en 2004. Ces critères actualisés de Bethesda sont plus sensibles que les critères d'Amsterdam II, permettant l'identification de certaines formes cliniques atypiques, environ 10 à 15 % des cas de syndrome de Lynch, qui ne répondent pas aux critères d'Amsterdam II. Cependant, il est estimé qu'aujourd'hui 28% des patients porteurs d'une mutation des gènes MMR restent non identifiés, reflétant la possible hétérogénéité de ce syndrome. C'est pourquoi ces recommandations pour la sélection clinique des cas évocateurs de syndrome de Lynch (Tableau 2) s'appuient aujourd'hui sur des analyses moléculaires et l'IHC pour mettre en évidence des signatures du syndrome de Lynch dans les tumeurs.

e. Importance du diagnostic moléculaire du syndrome de Lynch

Le diagnostic moléculaire du syndrome de Lynch est assuré par l'identification d'une mutation constitutionnelle délétère ou d'une épimutation dans l'un des gènes MMR (pour revue : Lynch *et al.*, 2015) en cohérence avec les caractéristiques tumorales observées, notamment l'analyse de l'instabilité microsatellitaire et l'étude de l'expression des protéines MMR dans les tumeurs. L'identification de l'altération causale est essentielle d'un point de vue diagnostique, pronostique et thérapeutique. En effet, l'identification d'une mutation délétère dans l'un des gènes MMR permet une meilleure prise en charge du malade et un suivi médical approprié des apparentés. Elle permet aux cliniciens de proposer des examens de surveillance aux individus porteurs de la mutation permettant un dépistage précoce, tandis qu'elle permet de lever une surveillance inutile, chère et chronophage, chez les individus non porteurs (qui ont un risque de développer un CRC équivalent à celui retrouvé dans la population générale). Ces examens correspondent à des coloscopies régulières (tous les 1-2 ans), complétées pour les femmes par des hystéroscopies, des échographies utérines et des biopsies utérines (pour revue : Vasen and de Vos

Tot Nederveen Cappel, 2013). Afin de diminuer le risque d'atteinte d'un cancer de l'endomètre ou de l'ovaire, une hystérectomie et une salpingo-oophorectomie (ablation chirurgicale simultanée des ovaires et des trompes de Fallope) prophylactiques peuvent être proposées aux femmes ménopausées.

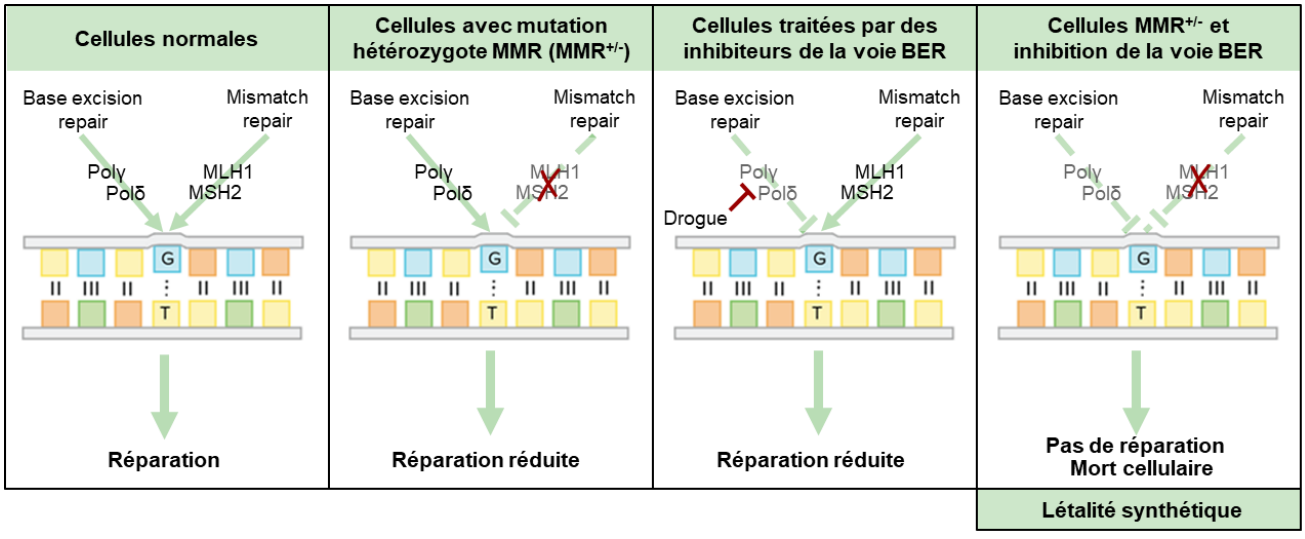
D'un point de vue pronostique, il a été montré que les CCR de phénotype MSI, en l'occurrence ceux associés au syndrome de Lynch, ont un meilleur pronostic à stade égal que ceux de phénotype MSS. Il a été estimé que le taux de survie à 5 ans chez les patients atteint d'un syndrome de Lynch est d'environ 60%, comparativement au 40-50% estimé dans les cancers sporadiques. En effet, du fait de l'instabilité microsatellitaire, il y a accumulation de mutations *frameshift* au niveau des séquences répétées de certaines régions codantes. Celles-ci génèrent alors des néopeptides et donc des néoantigènes qui vont être reconnus par les cellules lymphocytaires $CD8^+$ T (Maby *et al.*, 2016). Ainsi, les CRC à MSI ont une densité en lymphocytes infiltrant la tumeur (TIL, *tumor infiltrating lymphocytes*) plus importante que les autres cancers, ce qui permet une meilleure réponse immunitaire de l'organisme. Cependant, malgré la grande densité de TILs, ces tumeurs ne sont généralement par éliminées par le système immunitaire étant donné l'expression de grande quantité de molécules immunosupressives, appelées *immune checkpoints*, sécrétées par les cellules tumorales déficientes en activité MMR et le microenvironnement, en particulier PD-1, PD-L1, CTLA-4, LAG-3 et IDO (Llosa *et al.*, 2015; pour revue : Westdorp *et al.*, 2016). Ces données laissent suggérer que des stratégies d'immunothérapies personnalisées pourraient être proposées à ces patients (pour revues : Basile *et al.*, 2017; Bever and Le, 2017). Il pourrait s'agir de stratégies basées sur (i) l'inhibition de l'activité immunosupressive des *immune checkpoint* avec le développement d'anticorps monoclonaux dirigés contre ces molécules, en particulier PD-1 (pembrolizumab, nivolumab) et PD-L1 (atezolizumab), actuellement en essais cliniques, (ii) la vaccination contre le cancer, consistant à stimuler la capacité du système immunitaire à reconnaitre et détruire les antigènes tumoraux spécifiquement exprimés à la surface des cellules cancéreuses ou (iii) le transfert adoptif de lymphocytes T cytotoxiques (CTL, *cytotoxic T lymphocytes*) autologues spécifiques de la tumeur, consistant à stimuler *in vitro*, via par exemple des cellules présentatrices d'antigènes artificielle (CPA), les propres CTLs des patients contre des néopeptides immunogéniques dérivés des mutations *frameshift* spécifiques des tumeurs MSI (Latouche and Sadelain, 2000; Maby *et al.*, 2016).

L'identification de la mutation à l'origine de la maladie peut également contribuer à éviter l'administration de traitements anti-cancéreux inefficaces ou l'apparition de phénomène de résistance à certains traitements anti-cancéreux classiquement utilisés, tels que les agents méthylants, les sels de platine et les fluoropyrimidines (pour revue : Guillotin and Martin, 2014). En particulier, de nombreuses études ont révélé l'inefficacité du 5-FU (5-fluorouracile), anti métabolite analogue de la pyrimidine utilisé dans le traitement de certains cancers (Jover *et al.*, 2009; Ribic *et al.*, 2003; Sargent *et al.*, 2010). Il doit être métabolisé dans la cellule pour être actif. En effet, le 5-FU est transformé au sein des cellules cancéreuses, via les mêmes réactions enzymatiques que les nucléotides normaux, en différents métabolites cytotoxiques qui sont incorporés dans l'ADN et l'ARN induisant *in fine* l'arrêt du cycle cellulaire et l'apoptose. Plus précisément, il agit principalement sur la synthèse d'ADN sous forme de 5-FdUMP (5 fluorodésoxy-uracile monophosphate) en bloquant l'activité de la thimidylate synthase, qui catalyse la conversion du deoxyuracil monophosphate (dUMP) en deoxythymidine monophosphate (dTMP). Ceci conduit à une carence en thymine, utilisée pour la synthèse de l'ADN et par conséquent une substitution par l'uracile et la 5-FU. De plus, un autre de ses métabolites, le 5-FUTP, a la capacité de s'incorporer dans les divers types d'ARN, ce qui conduit à une transcription erronée. Dans le cas des cellules cancéreuses déficientes en activité MMR (tumeurs de phénotypes MSI), il a été démontré, d'abord *in vitro* sur des lignées issues de tumeurs CCR, que le traitement par le 5-fluorouracile était moins efficace contre les cellules cancéreuses proficientes en activité MMR (tumeurs de phénotype MSS) (Carethers *et al.*, 1999). Ces résultats ont ensuite été confirmés par des études cliniques montrant que les cellules tumorales MSI présentaient une moindre sensibilité à la chimiothérapie par le 5-FU, ce qui se traduit par aucune amélioration de la survie pour les CCR MSI traités par 5-FU (Jover *et al.*, 2009; Ribic *et al.*, 2003; Sargent *et al.*, 2010). Aujourd'hui, certaines données préliminaires suggèrent que l'adjonction d'oxaliplatine au 5-FU pourrait rétablir le bénéfice de la chimiothérapie par 5-FU chez les patients avec une tumeur MSI. Le phénotype MSI étant un facteur de résistance au 5-FU, la détermination du statut microsatellitaire est par conséquent indispensable pour la sélection d'une chimiothérapie adaptée.

Les patients évocateurs d'un syndrome de Lynch pour lesquels une mutation dans l'un des gènes MMR a été identifiée pourront éventuellement disposer de nouvelles thérapies ciblées. Récemment, de nouvelles approches thérapeutiques basées sur le concept de létalité synthétique, définie comme la mort cellulaire obtenue par la perte de fonction concomitante de deux voies de

réparation complémentaires, qui individuellement ne sont pas létales, ont été proposées pour le traitement des tumeurs associées au syndrome de Lynch (Figure 8 ; pour revue : McLornan *et al.*, 2014). Ces nouvelles approches ciblent notamment les ADN polymérases impliquées dans la voie de réparation par excision de base (Base, *base excision repair*), synthétiquement létal pour les cellules déficientes en activité MMR, par perte de PSH2 ou MLH1, respectivement (pour revue: Guillotin and Martin, 2014). En effet, l'ADN polymérase ß impliqué dans la voie BER est liée à *MSH2* tandis que l'ADN polymérase mitochondrial γ est liée à *MLH1* de sorte que l'inhibition de la polymérase induit une accumulation de lésions oxydatives (8-oxoG) au niveau de l'ADN nucléaire ou mitochondrial, spécifiquement dans les cellules tumorales déficientes en activité MMR (pour revue : Martin *et al.*, 2010). Il a ainsi été montré que l'utilisation du méthotrexate (MTX), un antifolique, conduit à l'accumulation de lésions oxydatives, en particulier la 8-oxoguanine, dans les cellules déficientes en *MSH2* spécifiquement du fait de leur difficulté à réparer ces dommage (Martin *et al.*, 2009). Cette étude suggère que le système MMR joue un rôle de *back-up* dans la réparation des dommages oxydatifs et que la perte des deux voies de réparation des dommages oxydatifs, BER et MMR, entraîne une surcharge de lésions non réparées qui affecte la viabilité cellulaire. Un essai clinique de phase II sur les effets cytotoxiques du méthotrexate est actuellement en cours (pour revue : Guillotin and Martin, 2014).

Figure 8 : Mécanisme de la mort cellulaire spécifique des cellules déficientes en système MMR provoquée par la létalité synthétique, elle-même induite par l'inhibition des polymérases δ et γ impliquées dans la voie de réparation par excision de base (adapté de Iglehart and Silver, 2009). BER, *base excision repair* ; MMR, *mismatch repair*.



3) La prédisposition génétique aux cancers du sein et de l'ovaire

Les cancers du sein et de l'ovaire constituent, respectivement, la première et la huitième cause de mortalité féminine par cancer dans le monde, représentant plus d'un tiers des nouveaux cas de cancers diagnostiqués annuellement chez les femmes (Binder-Foucard *et al.*, 2014). On estime aujourd'hui qu'environ une femme sur neuf développera un cancer du sein au cours de sa vie. Il s'agit du cancer le plus mortel au sein de la population féminine. Alors que la majorité des cancers du sein et de l'ovaire sont sporadiques, résultant de mutations somatiques acquises au cours de la vie de l'individu, 5 à 10% des cancers du sein et de l'ovaire sont dus à une prédisposition génétique héréditaire (pour revue : Kobayashi *et al.*, 2013). On parle alors de syndrome seins-ovaires ou syndrome HBOC. Ce syndrome est caractérisé soit par la présence d'une histoire familiale avec un excès d'apparentés diagnostiqués pour un cancer du sein et/ou de l'ovaire avant la ménopause dans une même branche parentale, la précocité de survenue du cancer du sein (avant 40 ans pour le cancer du sein et avant 70 ans pour le cancer de l'ovaire), des formes de cancers du sein bilatérale, l'association chez un même individu d'un cancer du sein et de l'ovaire et la présence d'un cancer du sein chez un apparenté masculin.

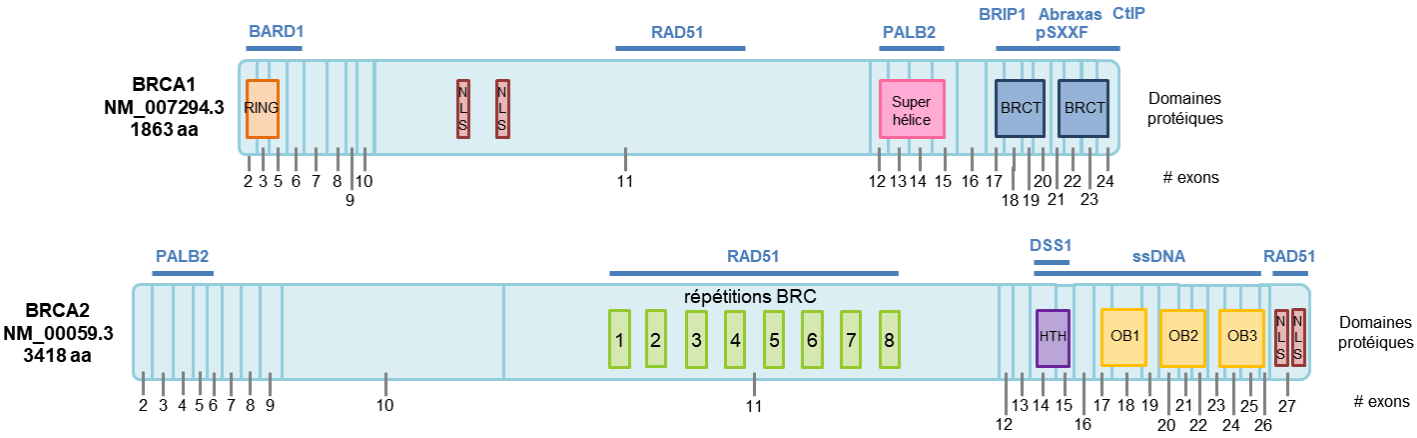
Les premiers cas de cancers du sein héréditaires ont été décrits par Paul Broca, chirurgien et pathologiste français, dans son « traité des tumeurs » (Paris, 1866) dans lequel il rapportait le cas d'une patiente suivie par l'un de ses confrères qui développa, très jeune, un cancer du sein et de sa famille dans laquelle 15 cas de décès par cancers, dont 10 par cancer du sein, avaient été rapportés sur quatre générations successives (pour revue : Rahman, 2014). Il s'agit ici d'un des premiers rapports décrivant la prédisposition héréditaire au cancer. Environ cent ans plus tard, Henry Lynch rapporte également une agrégation de cas de cancers du sein et de l'ovaire chez certaines familles qui apparaissent à un âge précoce (Lynch and Krush, 1971; Lynch *et al.*, 1974, 1976). L'analyse de ces familles a permis de mettre en évidence l'existence d'une prédisposition héréditaire aux cancers du sein et de l'ovaire, transmise selon un mode autosomique dominant et de pénétrance élevée (Claus *et al.*, 1991; Lynch *et al.*, 2013).

a. Les gènes BRCA, gènes majeurs de prédisposition au syndrome seins-ovaires

Le syndrome seins-ovaires est une prédisposition génétique héréditaire qui obéit à un déterminisme mendélien selon un mode de transmission autosomique dominant, avec une pénétrance forte mais incomplète (environ 90%). Deux gènes majeurs de prédisposition ont été identifiés : *BRCA1* (*BReast Cancer 1* ; OMIM #113705), gène de 80 Kb localisé au niveau de la région 17q21.3 et composé de 24 exons dont 22 codent pour une protéine de 1863 acides aminés et *BRCA2* (*BReast Cancer 2* ; OMIM #600185), gène de 84 Kb localisé au niveau de la région 13q12-13 et composé de 27 exons codant pour une protéine de 3418 acides aminés (Figure 9 ; Hall *et al.*, 1990; Miki *et al.*, 1994; Wooster *et al.*, 1995). Les gènes *BRCA1* et *BRCA2* partagent certaines caractéristiques, en particulier la présence d'un premier exon non codant et d'un exon central, l'exon 11, qui code pour plus de 50% de la protéine. Pourtant, les risques cumulés de développer un cancer du sein ou de l'ovaire à 70 ans varient en fonction du gène, tel que ce risque a été estimé à 60% et 55% de développer un cancer du sein et 59% et 16.5% pour un cancer de l'ovaire lorsque des mutations pathogènes *BRCA1* et *BRCA2* ont été identifiées, respectivement (Mavaddat *et al.*, 2013). Les gènes *BRCA1* et *BRCA2*, appelés dans leur ensemble gènes BRCA, codent pour des protéines notamment connues pour leur implication dans la réparation des lésions de l'ADN de type cassures double brin d'ADN (DSB, *double-strand break*) par recombinaison homologue (HR, *recombination homologue*) (Section I.3c). Mais ces protéines, aux fonctions multiples, sont plus largement impliquées dans de nombreux processus cellulaires.

La protéine BRCA2 est très largement impliquée dans le maintien de la stabilité du génome (Figure 10 ; pour revues : Fradet-Turcotte *et al.*, 2016; Martinez *et al.*, 2015). En effet, il a été mis en évidence que la déficience en BRCA2 par *knock-out* est létale au stade embryonnaire chez la souris causant une hypersensibilité aux irradiations et des réarrangements chromosomiques, suggérant ainsi que BRCA2 joue un rôle dans le maintien de la stabilité du génome (Sharan *et al.*, 1997). La protéine BRCA2 est notamment connue pour son implication dans la réparation des lésions de l'ADN, en particulier de type DSBs par RH (Section I.3c) et de type liaisons intra-brins (ICLs, *instrastrand crosslinks*) par la voie de l'anémie de Fanconi. Ces dernières correspondent à des lésions toxiques entre un des deux brins d'ADN et une autre molécule dite agent pontant qui empêchent la réplication et la transcription en inhibant la séparation des brins (Deans and West, 2011). Ce type de dommages peuvent être notamment induits par certaines chimiothérapies

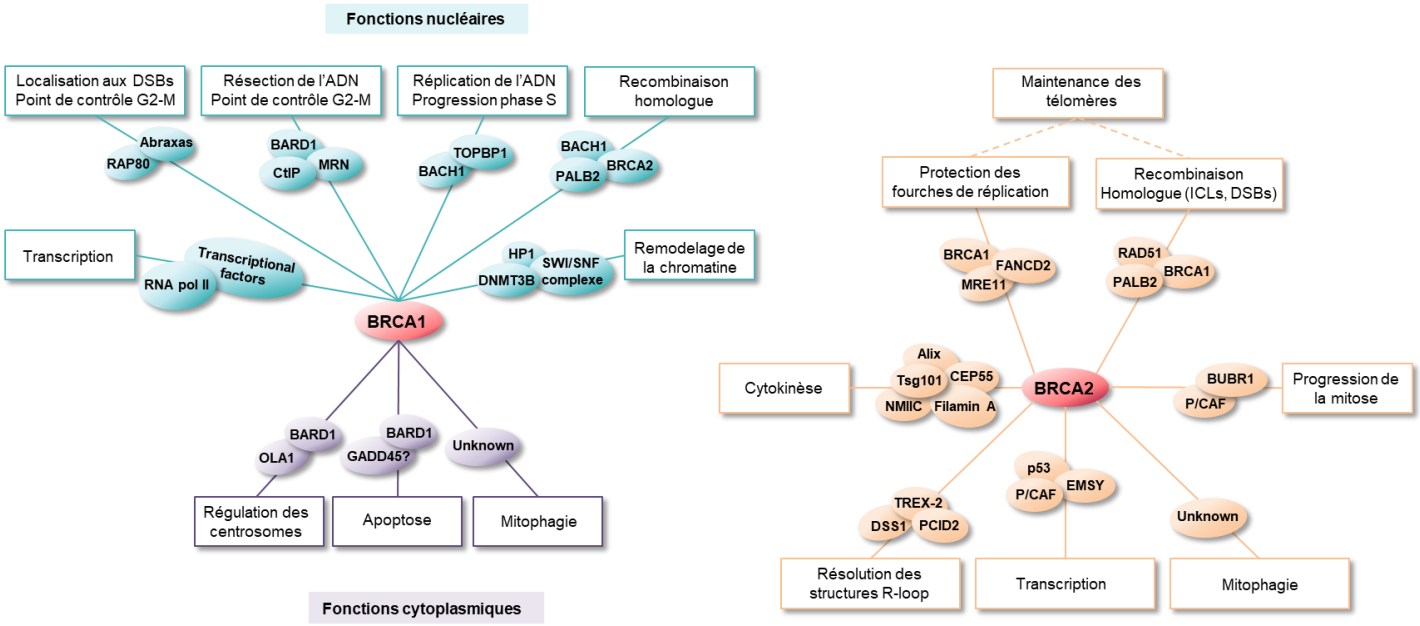
Figure 9 : Représentation des protéines BRCA incluant les domaines fonctionnels et leurs correspondances avec les exons et les interacteurs protéiques (adapté de Liu *et al.*, 2012 et Roy *et al.*, 2012). BRCT, *BRCA1 C-terminal domain* ; ssDNA, *single strand DNA* ; HTH, *Helix-turn-helix ;* NLS, *nuclear localization signal* ; OB, oligonucleotide binding; pSXXF, phosphorylated sequence motif SXXF.



telles que la cisplatine et la mitomycine C (Deans and West, 2011). De plus, BRCA2 est impliquée dans la maintenance de l'intégrité des télomères, en facilitant leur réplication et leur protection. En effet, il a été montré que des tumeurs déficientes en BRCA2, mais pas en BRCA1, montraient un raccourcissement et une fragilité des télomères et une augmentation des échanges de télomères entre les chromatides sœurs *(Badie et al., 2010; Min et al.*, 2012). BRCA2, via le recrutement de RAD51, est aussi largement impliquée dans la protection, contre la résection excessive par les nucléases MRE11 et DNA2, des fourches de réplication. La protection de ces fourches de réplication contre une résection excessive est importante pour maintenir la stabilité du génome étant donné qu'elle peut conduire à une instabilité génomique dans les cellules cancéreuses déficientes en BRCA1/2 ou d'anémie de Fanconi. La progression de ces fourches peut être stoppée par des lésions de l'ADN (ICLs, altérations de bases, cassures simples brins), des structures secondaires (R-loops et G-quadruplexe), des éléments de répétitions ou la déplétion en certains nucléotides. D'ailleurs BRCA2 avec le complexe TREX-2, est aussi nécessaire pour l'élimination des boucles R (*R-loops*), hybrides ARN/ADN qui correspondent à des produits normaux issus de la transcription. Cependant, lorsqu'ils persistent, il peuvent engendrer le blocage des fourches de réplication (Aguilera and García-Muse, 2012). BRCA2 est aussi impliquée dans le contrôle de la transcription, en inhibiant l'activité transcriptionnelle de cibles telles que EMSY, P/CAS et P53 (Milner *et al.*, 1997; Rajagopalan *et al.*, 2010). Par ailleurs, BRCA2 est aussi impliquée dans le

point de contrôle du cycle cellulaire G2/M, la mitose et la cytokinèse (Menzel *et al.*, 2011). Plus récemment, la protéine BRCA2 a été impliquée dans la mitophagie et la clairance des mitochondries endommagées, processus cytoplasmique qui cible spécifiquement les mitochondries endommagées vers une autophagie sélective (Sumpter *et al.*, 2016). L'ensemble des diverses fonctions de *BRCA2* décrites à ce jour font que cette protéine est ainsi considérée comme l'un des gardiens de la stabilité du génome en particulier lors de stress réplicatifs (pour revues : Fradet-Turcotte *et al.*, 2016; Martinez *et al.*, 2015).

<u>Figure 10</u> : Les multiples rôles des protéines BRCA1 et BRCA2 (adapté de Martinez *et al.*, 2015 et Takaoka and Miki, 2018).



Tout comme BRCA2, BRCA1 exerce de nombreuses fonctions contribuant toutes à préserver l'intégrité du génome (Figure 10 ; pour revues : Savage and Harkin, 2015; Takaoka and Miki, 2018). Cette protéine, pléiotrope, exerce la plupart de ces fonctions, par l'intermédiaire d'interactions avec un très grand nombre de protéines, plus d'une centaine, avec lesquelles elle interagit pour former des complexes protéiques (Roy *et al.*, 2011). Il a été rapporté que la déficience en BRCA1 par *knock-out* était létale au stade embryonnaire chez la souris, suggérant que BRCA1 était indispensable à la prolifération des cellules souches (Hakem *et al.*, 1996). Outre son implication dans la RH, BRCA1 participe à la régulation du cycle cellulaire en contrôlant de

multiples transitions du cycle cellulaire via un changement de partenaires au niveau de son domaine BRCT. Plusieurs complexes incluant BRCA1 ont été décrits comme régulateurs du cycle cellulaire, en particulier le complexe BRCA1–RAP80 activant le point de contrôle G2/M et le complexe BRCA1–BACH1 nécessaire lors du point de contrôle de la réplication pendant la phase S (Gong *et al.*, 2010; Yu and Chen, 2004). BRCA1 est également impliquée dans la modulation de l'expression des gènes. Tout d'abord, en formant de multiples complexes, BRCA1 exerce sur un certain nombre de ces cibles, un contrôle de l'expression génique au niveau transcriptionel : (i) le BRCT peut activer la transcription lorsqu'il est fusionné avec un domaine de liaison de l'ADN hétérologue et plusieurs mutations germinales identifiées chez des patients altèrent cette activité (Monteiro *et al.*, 1996), (ii) BRCA1 peut être associée à l'holoenzyme ARN polymérase II et de nombreux facteurs de transcription (Scully *et al.*, 1997), et (iii) BRCA1 peut favoriser la transcription de certains gènes rapporteurs lorsqu'il est surexprimé (Ouchi *et al.*, 1998; Zhang *et al.*, 1998). BRCA1 exerce également un contrôle de l'expression des gènes en participant au remodelage de la chromatine. Il a été suggéré que BRCA1 intervenait au niveau de la modification des histones afin de maintenir l'intégrité de l'hétérochromatine en régulant notamment l'ubiquitinisation de l'histone H2A (Zhu *et al.*, 2011). De plus, la protéine BRCA1 participerait au maintien de l'inactivation du chromosome X (Ganesan *et al.*, 2002). Il a également été attribué à BRCA1 plusieurs fonctions cytoplasmiques, notamment la mitophagie, le contrôle des centrosomes et l'apoptose. La protéine BRCA1 cytoplasmique est impliquée dans l'induction de la voie apoptotique via GADD45, d'une manière indépendante de la voie p53 (Fabbro *et al.*, 2002; Shao *et al.*, 1996). Elle régule également au niveau cytoplasmique le nombre de centrosomes durant les phase S tardives et G2/M (Ko *et al.*, 2006; Matsuzawa *et al.*, 2014; Starita *et al.*, 2004) pour prévenir une éventuelle hypertrophie des centrosomes et une aneuploïdie fréquemment retrouvée dans les cancers du sein. Et, récemment, la protéine BRCA1 a été décrite comme ayant un rôle dans les phénomènes d'autophagie sélective (pour revues : Savage and Harkin, 2015; Takaoka and Miki, 2018).

b. Aspects moléculaires du syndrome seins-ovaires

Dans la majorité des cas, le syndrome seins-ovaires est lié à une altération hétérozygote constitutionnelle identifiée sur l'un des gènes BRCA, soit *BRCA1* ou *BRCA2*. Alors que le risque de développer un cancer du sein chez les femmes est estimé à 12% dans la population générale, le

risque cumulé d'atteinte d'un cancer du sein ou de l'ovaire à 70 ans chez les patientes porteuses d'une mutation dans *BRCA1* ou *BRCA2* est de 60 et 59% et de 55 et 16,5%, respectivement (Mavaddat *et al.*, 2013; pour revue : Smith, 2012). La présence d'une mutation délétère dans un de ces deux gènes augmente ainsi considérablement le risque de développer un cancer du sein ou de l'ovaire. Il est à noter que les cancers héréditaires de l'ovaire sont le plus souvent associés à *BRCA1* comparativement à *BRCA2*. Chez les hommes en revanche, c'est *BRCA2* qui est plus sévèrement associé à la prédisposition au cancer du sein. En effet, les individus masculins porteurs d'une mutation *BRCA1* ou *BRCA2* présentent aussi un risque plus élevé de développer un cancer du sein, risque à l'âge de 80 ans de 1,2% et 6,8%, respectivement contre 0,12% dans la population masculine générale (Tai *et al.*, 2007; pour revue : Smith, 2012). De plus, les mutations au niveau des gènes BRCA, en particulier *BRCA2*, augmentent le risque de cancer de la prostate, de 0.1% dans la population générale à 20% chez les hommes porteurs d'une mutation dans *BRCA2* (pour revue : Smith, 2012). D'autres cancers, appartenant au spectre tumoral du syndrome seins-ovaires, parmi lesquels les mélanomes ou les cancers du pancréas (de 0.5% dans la population générale à 10% chez les individus porteurs d'une mutation dans *BRCA2*) sont également associés à des mutations dans les gènes *BRCA1* et *BRCA2* (Tableau 4 ; pour revue : Smith, 2012).

Tableau 4 : Risque de cancers associés à la présence d'une mutation dans un gène BRCA (d'après Roy *et al.*, 2012)**.** CLL, chronic lymphocytic leukemia ; AML, acute myeloid leukemia.

Type de cancers	Mutations BRCA1	Mutations BRCA2
Sein	70-80% de risque	50-60% de risque
Ovaire	50% de risque	30% de risque
Prostate	Risque augmenté chez les Juifs ashkénazes porteurs d'une mutation fondatrice	20 fois plus de risque
Pancréas	Anecdotique Uniquement des études de cas	10 fois plus de risque
Estomac	Non reportés	Rapports limités
Autres	Non reportés	Cerveau, médulloblastomes, pharynx, CLL et AML
Trompes de Fallopes	Observées mais rares	Rares

Les variations détectées dans les gènes BRCA comprennent une proportion importante de réarrangements génomiques de grande taille. Ces derniers constituent 8-15% des mutations délétères des gènes BRCA associées au syndrome seins-ovaires (pour revue : Kobayashi *et al.*, 2013). En effet, les gènes BRCA contiennent une très grande densité d'éléments répétés, notamment des séquences Alu, représentant environ 45% des séquences génomiques de ces gènes (pour revue : Welcsh and King, 2001). Ces séquences Alu sont particulièrement associées à des réarrangements génomiques à l'origine d'un syndrome seins-ovaires (Nordling *et al.*, 1998; Unger *et al.*, 2000). Les mutations détectées peuvent également correspondre à des mutations ponctuelles de type indels, non-sens voire même d'épissage entrainant l'apparition d'un codon stop prématuré (PTC, *premature terminaison codon*) à l'origine d'une protéine tronquée probablement non fonctionnelle. A cela s'ajoute les variations ponctuelles au niveau de la région codante, de type faux-sens, survenant dans les domaines fonctionnels indispensables à l'activité de la protéine, en particulier au niveau des domaines de liaison à PALB2 et à l'ADN pour *BRCA2* (Biswas *et al.*, 2012; Guidugli *et al.*, 2014, 2018) et au niveau du domaine BRCT (*BRCA1* C-*terminal*) et RING (*really interesting new gene*) pour *BRCA1* (Bouwman *et al.*, 2013).

Pourtant, l'ensemble des gènes BRCA n'expliquent à eux seuls, qu'une faible proportion (entre 10 et 25%) des syndromes seins-ovaires (Kast *et al.*, 2016). Cette hérédité manquante peut s'expliquer aujourd'hui par l'implication d'autres gènes de prédisposition. En effet, l'évolution des connaissances sur le syndrome seins-ovaires a permis récemment l'identification d'un troisième gène de prédisposition : le gène *PALB2*, qui confère un risque similaire à *BRCA2* de développer un cancer du sein (Antoniou *et al.*, 2014). D'autres gènes seraient également impliqués dans le déterminisme génétique des cancers du sein et/ou de l'ovaire, notamment un panel de 17 autres gènes constitués des gènes *ATM*, *BARD1*, *BRIP1*, *CDH1*, *CHEK2*, *NBN*, *PTEN*, les paralogues du gène *RAD51*, *RAD51B* , *RAD51C* et *RAD51D*, *STK11* et *TP53*, les gènes de l'anémie de Fanconi (FA, *fanconi anemia*) et enfin les gènes MMR (*MLH1*, *MSH2*, *MSH6*, *PMS2* et *EPCAM*) (Bubien *et al.*, 2013; Byrnes *et al.*, 2008; Giardiello *et al.*, 2000; Golmard *et al.*, 2013; Hilbers *et al.*, 2012; Loveday *et al.*, 2012; Meindl *et al.*, 2010; Park *et al.*, 2012; Pharoah *et al.*, 2001; Walsh *et al.*, 2006; pour revues : Kobayashi *et al.*, 2013; Smith, 2012). Mais la contribution respective de ces gènes dans le déterminisme du syndrome seins-ovaires ainsi que la pénétrance de ces mutations restent à caractériser.

Etant impliqués dans l'une des voies de réparation des dommages de l'ADN, les gènes BRCA sont considérés, au même titre que les gènes MMR, comme des gènes suppresseurs de tumeurs garantissant le maintien de la stabilité et de l'intégrité du génome. Ces gènes répondent donc également au modèle à deux coups de Knudson où une inactivation des deux allèles est nécessaire pour conduire à l'apparition d'une tumeur, l'une étant héritée dès la naissance et l'autre acquise au cours de la vie (pour revue : Friedberg, 2003). Les cellules épithéliales mammaires, comportant une altération constitutionnelle de l'un des gènes BRCA, sont soumises en réponse aux estrogènes à une rapide prolifération durant la puberté et ce jusque la ménopause (pour revue : Welcsh and King, 2001). Il est très probable que l'augmentation spectaculaire du taux de réplication de ces cellules affecte leur capacité de réparation de l'ADN. De plus, ces gènes comportant une très forte densité d'éléments répétés, ils sont notamment soumis à des réarrangements de grande taille à l'origine de pertes d'hétérozygotie (LOH, *loss of heterozygosity*) qui entrainent l'inactivation du second allèle. Suite à l'inactivation des 2 allèles, le système de RH devient complètement non fonctionnel, aboutissant à une accumulation des erreurs lors de la réplication de l'ADN qui ne pourront pas être réparées lors du cycle de réplication suivante, dans les cellules déficientes en BRCA, ce qui va conduire à l'activation de la voie de réponse des dommages de l'ADN (DDR, *DNA damage response*), des points de contrôles du cycle cellulaire et *in fine* à la mort cellulaire programmée. Mais, dans l'épithélium mammaire où les cellules prolifèrent rapidement, certaines des cellules déficientes dans leur activité de réparation peuvent échapper à ces systèmes de surveillance et donc à l'apoptose. En effet, l'instabilité génomique va conduire à l'accumulation de mutations additionnelles qui vont donner à ces cellules, via les modifications de phénotype qu'elles provoquent, un avantage sélectif décrit comme les « *hallmarks of cancer* » par Hanahan et Weinberg, qui se traduit par une amélioration de la survie et de la reproduction (pour revues : Hanahan and Weinberg, 2000, 2011). On parle alors de mutations *drivers*. D'autres mutations, dites *passenger* peuvent également être acquises par ces cellules, mais celles-ci sont supposées ne pas avoir d'impact sur la tumorigenèse.

Alternativement, il a été proposé un autre modèle d'haploinsuffisance de *BRCA1*, à la suite de plusieurs découvertes récentes qui laissent suggérer que la présence d'une mutation délétère sur un seul des 2 allèles serait suffisante à elle seule pour altérer les cellules épithéliales du sein ou de l'ovaire (Sedic and Kuperwasser, 2016). Tout d'abord, il a été suggéré que la présence d'une mutation délétère sur une des deux allèles serait responsable de la diminution de l'expression de la

protéine (~ 50%) (Baldeyron *et al.*, 2002) et que certaines protéines BRCA1 tronquées dans leur partie C-terminale pourrait abroger certaines fonctions de la protéine BRCA1 WT, notamment la chimiosensibilité, la susceptibilité à l'apoptose et l'inhibition de l'activité de transcription des récepteurs aux œstrogènes (Fan *et al.*, 2001). De plus, il a été montré que les cellules porteuses d'une seule mutation de *BRCA1* présenteraient des défauts dans les réponses de réparation des dommages à l'ADN ainsi qu'une instabilité génomique et une sensibilité au stress génotoxique accrues et une augmentation du taux d'hyper recombinaisons spontanées comparativement à des cellules WT (Baldeyron *et al.*, 2002; Cousineau and Belmaaza, 2007; Konishi *et al.*, 2011; Rennstam *et al.*, 2010). L'ensemble de ces observations révèlent l'existence possible d'un modèle d'haploinsuffisance conditionnelle de *BRCA1* dans cellules hétérozygotes, c'est-à-dire que les niveaux basals de *BRCA1* dans des cellules haplo-insuffisantes sont suffisants pour assurer la plupart des fonctions de BRCA1, mais le *pool* disponible de cette protéine est en effet limité. Ainsi, sans inactivation du second allèle, des défauts deviennent apparents lorsque les cellules sont stimulées dans des conditions qui nécessitent une l'intervention de BRCA1 (pour revue : Sedic and Kuperwasser, 2016).

Une déficience constitutionnelle en BRCA peut également être observée chez des patients atteints d'anémie de Fanconi (Kee and D'Andrea, 2012; de Winter and Joenje, 2009). Ce syndrome héréditaire, de transmission récessive, autosomique ou alors liée à l'X, se caractérise par un dysfonctionnement de la moelle osseuse avec pour conséquence un déficit de production de cellules sanguines. Il est dû à des variations constitutionnelles bi-alléliques retrouvées à l'état homozygote ou hétérozygote composite dans l'un des 16 gènes identifiés comme étant responsable de l'AF, dont *BRCA2* (FANCD2) (Howlett *et al.*, 2002), *PALB2* (FANCN) (Reid *et al.*, 2007; Xia *et al.*, 2007) et plus récemment *BRCA1* (Garcia-Higuera *et al.*, 2001; Sawyer *et al.*, 2015), gènes impliqués dans la réparation de l'ADN notamment au niveau de la voie de réparation de Fanconi.

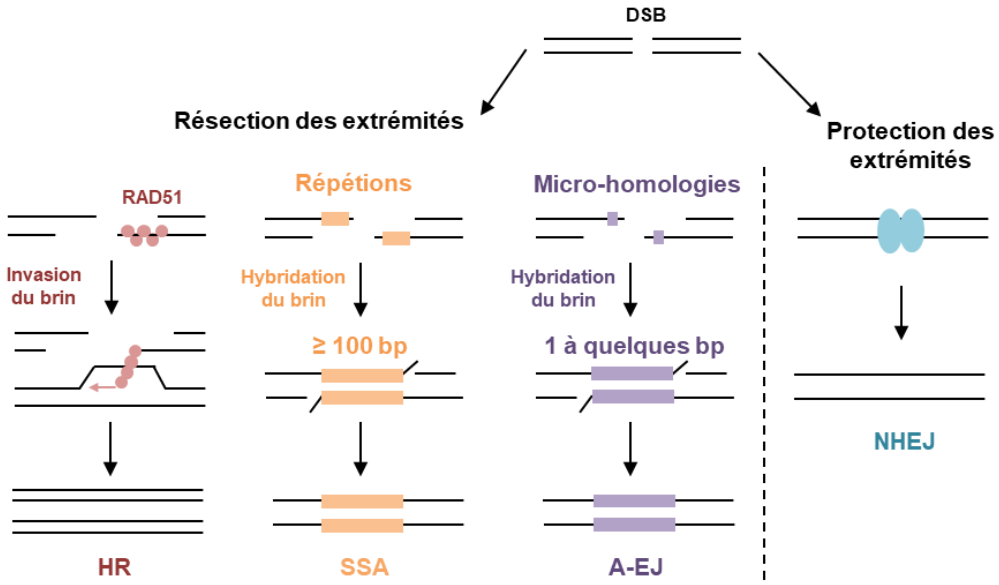
c. Le système de réparation des cassures double brin par recombinaison homologue

Le système de RH est, au même titre que le système MMR, un acteur majeur du maintien de l'intégrité de l'information génétique. Mais contrairement au système MMR dédié à la réparation des erreurs réplicatives de type mésappariements, ce système est voué à la réparation des DSBs survenant suite l'exposition des cellules à des agents génotoxiques physiques (radiations

ionisantes) ou chimiques (espèces réactives de l'oxygène, composés alkylants, chimiothérapies de type cisplatine). On parle alors de causes exogènes. Les cellules elles-mêmes peuvent être à l'origine des cassures double brin. Il s'agit dans ce cas de causes endogènes, qui correspondent en particulier (i) aux espèces réactives de l'oxygène (les radicaux libres) générées par le métabolisme même de ces cellules, (ii) à l'arrivée d'une fourche de réplication sur une cassure simple-brin ou sur une altération de l'ADN qui empêche sa progression et (iii) à un stress mécanique infligé à l'ADN compacté en chromosome. Jusqu'à cinquante DSBs peuvent se former par cycle cellulaire (pour revue : Vilenchik and Knudson, 2003). Et celles-ci, considérées comme les lésions de l'ADN les plus cytotoxiques, peuvent être lourdes de conséquences pour la cellule. En effet, l'accumulation de cassures non réparées induit généralement un arrêt du cycle cellulaire et/ou la mort cellulaire par apoptose. D'ailleurs, une seule DSB non réparée suffit à induire la mort cellulaire (Bennett *et al.*, 1996; Sandell and Zakian, 1993). Quant aux cassures mal réparées, elles peuvent être à l'origine de remaniements chromosomiques favorisant potentiellement l'instabilité génomique et le processus de carcinogenèse (pour revue : Jeggo and Löbrich, 2007).

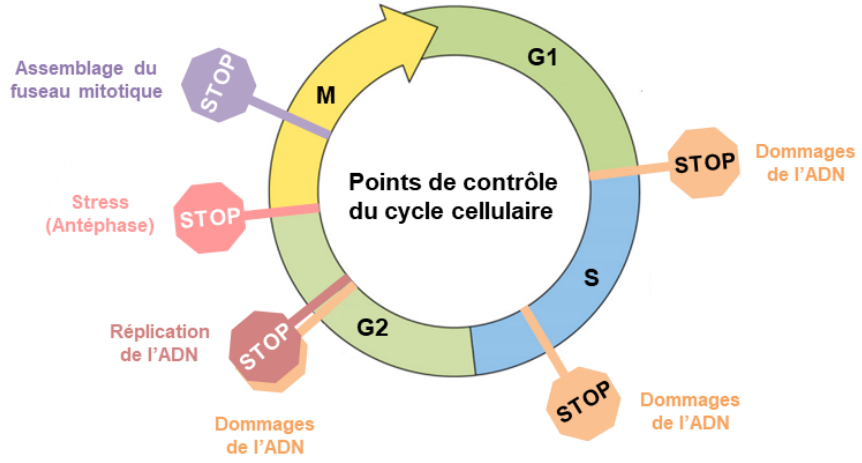
Deux principaux mécanismes peuvent procéder à la réparation de ces cassures, la réparation par jonction des extrémités non-homologues (NHEJ, *non homologous end joining*) ou par recombinaison homologue (Figure 11 ; pour revue : Aparicio *et al.*, 2014). Alors que le NHEJ consiste en la ligature directe des deux extrémités de la cassure, la RH utilise une séquence d'ADN homologue à celle endommagée (la chromatide sœur) pour recopier et reconstituer la région perdue lors de la cassure. Ainsi, la RH représente le mécanisme de réparation des DSBs le plus efficace, puisque le NHEJ génère fréquemment des erreurs pouvant entraîner une instabilité du génome. Mais ce mécanisme ne peut intervenir que lorsque l'ADN cellulaire est répliqué, lors des phases S et G2 du cycle cellulaire, contrairement au NHEJ qui peut opérer tout au long du cycle cellulaire. D'autres voies alternatives de réparation des DSBs existent, notamment une voie alternative de ligature d'extrémités non-homologues (A-EJ, *alternative end joining*) médiée par la micro-homologie, et une voie de protection des extrémités consistant à la ligation des extrémités situées de part et d'autres de la cassure, qui sont cependant source d'erreurs.

Figure 11 : Les voies de réparation des cassures double-brin de l'ADN (d'après Jasin and Rothstein, 2013). DSB, *double strand break* ; HR, *homologous recombinaison* ; SSA, *single-strand annealing* ; A-EN, *alternative end-joining* ; NHEJ, *non homologous end joining* ; pb, paires de bases.



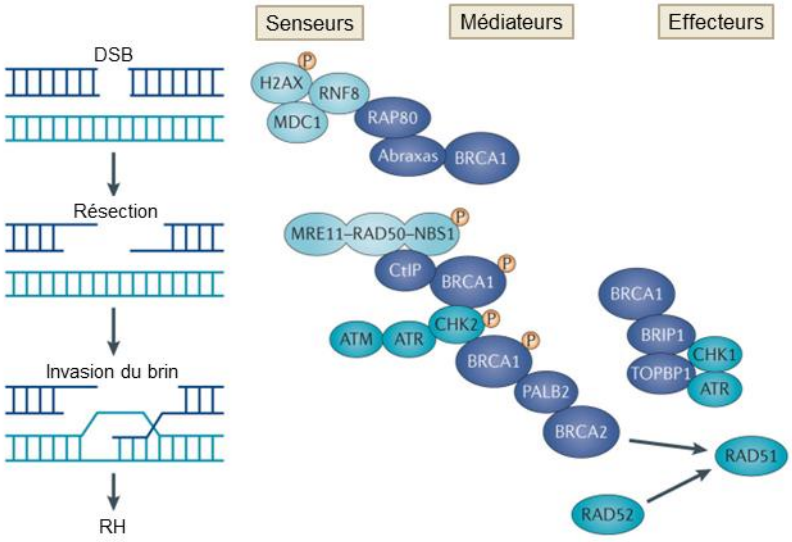
Ces mécanismes de réparation sont étroitement liés à un réseau de surveillance dépendant notamment des points de contrôle du cycle cellulaire (*checkpoints*). Lorsqu'ils sont activés, ces points de contrôles vont ralentir ou arrêter le cycle cellulaire et activer parallèlement les mécanismes de réparation. En effet, la progression dans le cycle cellulaire est contrôlée par des points de contrôle, défini comme « des voies de régulations biochimiques qui contrôlent la progression du cycle cellulaire pour éviter le début de certaines réactions avant que les précédentes ne soient terminées » (Sancar *et al.*, 2004). Six points de contrôle ont été identifiés (Figure 12). Il s'agit du point de contrôle de la transition G1/S, de l'intra-phase-S (ou intra-S), de la transition G2/M et de réplication, d'entrée en mitose (antéphase) et de l'assemblage du fuseau mitotique. Si les trois derniers répondent plus spécifiquement à un défaut au niveau de la réplication, à un stress avant l'entrée en mitose et un défaut d'attachement des chromosomes aux microtubules du fuseau, respectivement, les autres points de contrôle bloquent la progression du cycle cellulaire si l'ADN est endommagé, tout en favorisant la réparation des dommages de l'ADN pour empêcher la transmission des erreurs d'ADN aux cellules filles. C'est ce que l'on appelle la de la voie de réponse des dommages de l'ADN.

Figure 12 : Les points de contrôle du cycle cellulaire (adapté de Fei Chin and May Yeong, 2010).



Le système de réparation des DSBs fait intervenir 3 éléments : des « détecteurs » (*sensors*) des extrémités d'ADN cassées, des « effecteurs » (*effector*) pour exécuter la réparation, et des « médiateurs » (*signal transducers*) qui font le lien entre les « détecteurs » et les « effecteurs » (Figure 13 ; pour revue : Roy *et al.*, 2011). Dans le système de recombinaison homologue, les serine/thréonine kinases ATM et ATR (*ataxia-talangiectasia mutated and Rad3-related*), couplées à d'autres serine/thréonine kinases Chk2 et Chk1 (*checkpoint kinase 1/2*) formant les voies ATM-Chk2 et ATR-Chk1 interviennent dans la reconnaissance des DSBs et dans le blocage de la

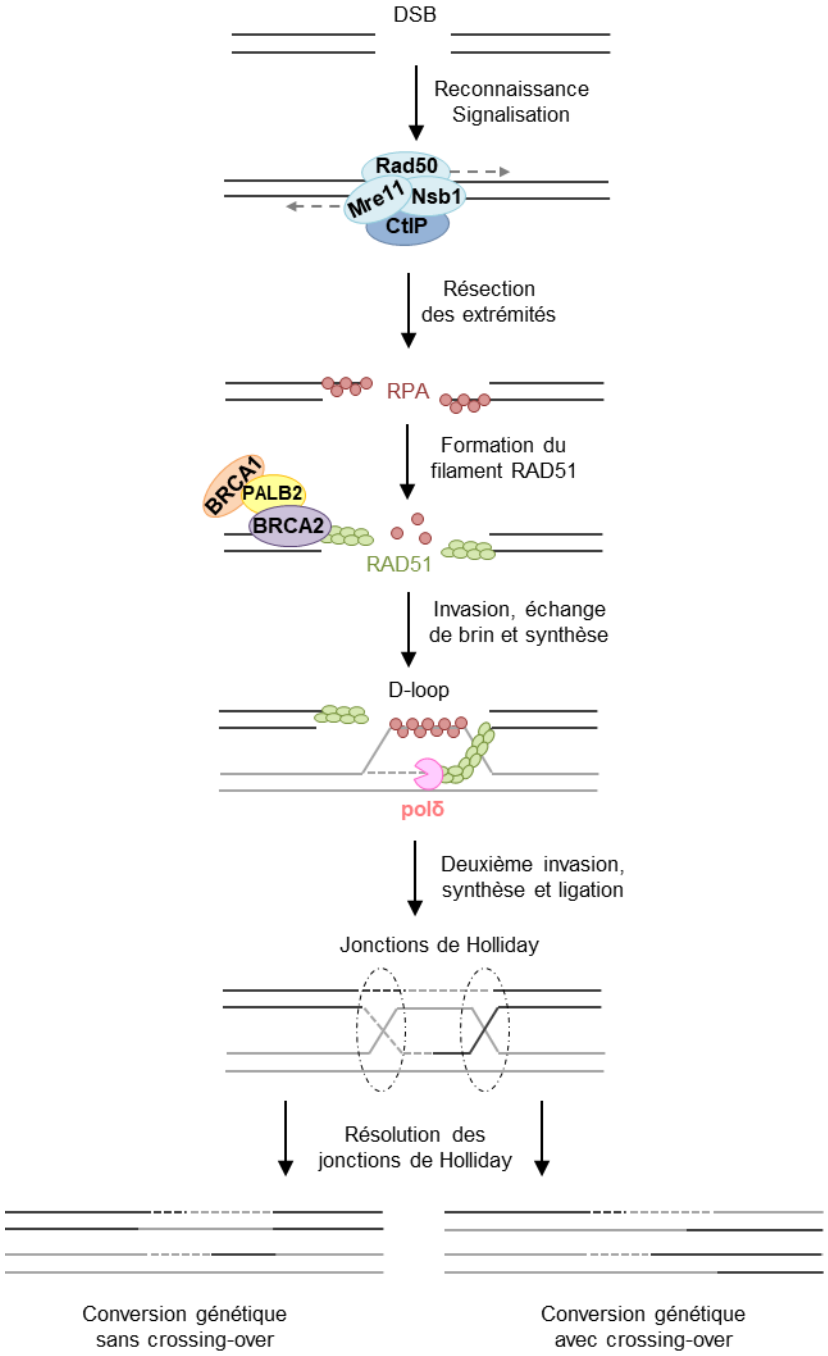
Figure 13 : Les principaux acteurs de la réponse aux dommages de l'ADN de type DSBs (d'après Roy *et al.*, 2011).



fourche de réplication, respectivement. Suite à la survenue de DSBs, le complexe MRN, formé des protéines Mre11, Rad50 et Nsb1, et la protéine CtIP sont recrutées au niveau des DSBs et initient la résection des extrémités de part et d'autre des DSB (Figure 14). Les extrémités simples brins sont alors stabilisées par la protéine RPA (*replication protein A*), ce qui permet alors le recrutement de protéines, parmi lesquelles BRCA1, PALB2, BRCA2, 53BP1, RAD51 et ses homologues. Sur les extrémités simple brin, la protéine RAD51 forme autour de l'ADN un filament nucléoprotéique ou filament présynaptique, favorisant la recherche d'une homologie avec la chromatide sœur puis l'envahissement de celle-ci par formation d'une D-loop (*displacement loop* ou boucle de déplacement), structure d'ADN comportant un ADN double-brin hétéroduplexe et un ADN simple-brin déplacé. Une fois la région homologue envahie, l'extrémité 3' du brin envahissant est allongée par l'ADN polymérase en utilisant l'ADN complémentaire du brin intact envahi comme matrice et les brèches simples brins vont finalement être refermées par la ligase. Il en résulte la formation d'une structure avec entrecroisement de brins homologues de deux chromatides sœurs, structure appelée jonctions de Holliday qui vont être par la suite résolues par coupure simultanée (pour revue : Jasin and Rothstein, 2013).

Les protéines BRCA1 et BRCA2 jouent toutes les deux un rôle central dans la recombinaison homologue, bien qu'elles exercent des fonctions différentes (Figure 15). La protéine BRCA1 a un rôle pléiotrope dans la réparation des DSBs : elle oriente la cellule vers la RH plutôt que vers le NHEJ, active les points de contrôle du cycle lorsque des altérations de l'ADN ont été détectées et participe au mécanisme de réparation homologue, à deux étapes différentes. En effet, BRCA1 intervient à un stade précoce, pour promouvoir la résection des cassures double brins afin de générer des extrémités non franches d'ADN simple brin et le recrutement de RAD51 au niveau de cet ADN simple brin et à un stade plus tardif pour promouvoir le recrutement de BRCA2 via PALB2. En effet, BRCA1 initie la résection en interagissant avec le facteur de résection CtIP préalablement phosphorylé en formant ainsi le complexe BRCA1-C (pour revue : Prakash *et al.*, 2015). Cette interaction favorise une coopération avec la nucléase MRN, qui colocalise avec BRCA1, pour catalyser la résection de l'ADN. De plus, BRCA1 antagonise 53BP1, un suppresseur de la résection. Il a été montré que contrairement aux cellules déficientes en BRCA1, les cellules déficientes en BRCA1/53BP1 sont proficientes pour la RH. BRCA1 assure également le recrutement de BRCA2 en utilisant la protéine PALB2 comme intermédiaire. En effet, il a été montré que la déplétion de BRCA1 inhibe la fixation de PALB2, BRCA2 et RAD51 ; la déplétion

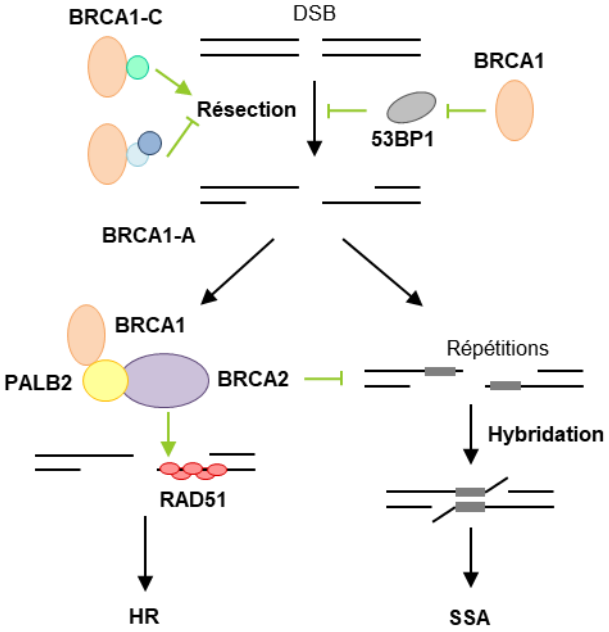
<u>Figure 14 :</u> Mécanisme de réparation des DSBs par recombinaison homologue dans les cellules eucaryotes.



de PALB2 diminue la fixation de BRCA2 et RAD51 ; et la déplétion de BRCA2 affecte seulement la fixation de RAD51 (pour revue : Prakash *et al.*, 2015). La protéine BRCA2, au même titre que PALB2 a un rôle plus direct et intervient spécifiquement dans le complexe protéique de

recombinaison homologue. Il s'agit de cofacteurs indispensables de la recombinaison homologue, appelés « médiateurs de la recombinaison homologue ». En effet, le recrutement de ces cofacteurs est nécessaire à (i) la formation du filament RAD51, en déplaçant la protéine RPA qui empêche la liaison de RAD51 sur l'ADN simple brin, (ii) la stabilisation des complexes RAD51-ADNss via l'inhibition de l'activité ATPase ADN-dépendante de RAD51 et (iii) la stimulation de l'activité recombinase de RAD51 permettant l'invasion (pour revue : Prakash *et al.*, 2015).

Figure 15 : Rôle des protéines BRCA1 et BRCA2 dans le mécanisme de recombinaison homologue (d'après Prakash *et al.*, 2015). DSB, *double strand break* ; HR, *homologous recombination* ; SSA, *single-strand annealing.*



d. Critères d'évaluation clinique évocateurs d'un syndrome seins-ovaires

Devant un contexte familial ou personnel évocateur d'une prédisposition héréditaire au cancer du sein ou de l'ovaire, il est maintenant possible de rechercher dans le génome de ces patients évocateurs, des mutations ponctuelles constitutionnelles hétérozygotes pathogènes ou des réarrangements génomiques dans l'un des gènes de prédisposition, *BRCA1*, *BRCA2* et *PALB2*, gènes clairement reconnus à très haut risque de cancer mais également dans d'autres gènes pour lesquels une augmentation de risque de cancers a été rapportée. En France, ce diagnostic

moléculaire repose notamment sur une analyse en panel de gènes, ciblant simultanément 13 gènes (*BRCA1*, *BRCA2*, *PALB2*, *TP53*, *CDH1*, *PTEN*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, *MSH6*, *PMS2* et *EPCAM*) pour lesquels le GGC-UNICANCER a conclu à une utilité clinique. Faute d'arguments suffisants, 7 autres gènes (*CHEK2*, *ATM*, *BARD1*, *BRIP1*, *NBN*, *RAD51B*, *STK11*) ne sont pour l'instant pas retenus pour cette analyse, sous réserve d'une éventuelle évolution du panel, compte tenu de la rapide évolution des connaissances et des résultats de l'étude TUMOSPEC visant à d'établir une estimation précise des risques de cancer associés aux gènes non retenus dans le panel.

Le diagnostic moléculaire survient suite à l'identification d'un certain nombre de signes évocateurs spécifiques du syndrome seins-ovaires permettant d'orienter le diagnostic moléculaire vers une prédisposition héréditaire liée à une altération de *BRCA1*, *BRCA2*, PALB2 (cancer du sein), RAD51 C et D (cancer de l'ovaire). Ces signes évocateurs correspondent principalement (i) des critères individuels tels qu'un âge de survenue très précoce d'un cancer du sein (avant l'âge de 36 ans) ou de l'ovaire (avant l'âge de 70 ans, tumeurs germinales et de type borderline mise à part), le sexe de l'individu (homme atteint d'un cancer du sein avant l'âge de 70 ans), l'ethnicité, notamment pour les femmes d'origine ashkénaze (la fréquence des mutations *BRCA* est 1/40 chez les Ashkénazes contre 1/500 dans la population générale), la bilatéralité des cancers du sein et la multiplicité des tumeurs (plusieurs cancers du sein ou de l'ovaire, association cancer du sein avec un cancer de l'ovaire/pancréas/prostate chez un même individu) et (ii) des critères familiaux traduisant une histoire familiale avec agrégation de cancers du sein et/ou de l'ovaire dans une famille, et plus précisément trois cas de cancer du sein au moins unis par des liens de parenté du 1er ou 2nd degré quels que soient les âges au diagnostic, deux cas de cancer du sein unis par des liens de parenté du 1er ou 2nd degré passant par un homme et dont dans au moins un cas, le diagnostic a été porté avant l'âge de 40 ans, ou une atteinte portée avant l'âge de 50 ans, l'autre avant l'âge de 70 ans, une femme atteinte d'un cancer du sein et au moins une apparentée unie par un lien de 1er ou 2nd degré passant par un homme, atteinte d'un cancer de l'ovaire, quels que soient les âges au diagnostic (et réciproquement).

Contrairement au syndrome de Lynch pour lequel une signature moléculaire dans la tumeur est parfaitement définie, il y a très peu de caractéristiques tumorales associées au syndrome seins-ovaires. En effet, le phénotype des tumeurs associées à ce syndrome est très similaire à celui observé pour les tumeurs sporadiques. Cependant, quelques sous-types de cancer du sein ont été

plus fortement retrouvés chez les patients évocateurs d'une prédisposition héréditaire, notamment les cancers du sein de type médullaire, forme très particulière de cancer du sein représentant moins de 1% des cancers du sein infiltrant, ou triples négatifs, caractérisés par l'absence de récepteurs hormonaux aux œstrogènes (RO-) et à la progestérone (RP-) et l'absence de surexpression du facteur de croissance HER-2 ou ERBb2 (HER2-) (Phuah *et al.*, 2012; Southey *et al.*, 2011; Spurdle *et al.*, 2014).

e. Importance du diagnostic moléculaire du syndrome seins-ovaires

Etant donné qu'une mutation constitutionnelle délétère dans l'un des gènes de prédisposition au syndrome seins-ovaires augmente le risque de développer un cancer du sein et/ou de l'ovaire chez les patients porteurs, l'identification, chez ces patients, d'une mutation causale dans l'un de ces gènes est donc essentielle pour l'optimisation de la prise en charge des patients et de leurs apparentés. Cette optimisation de la prise en charge consiste d'abord, pour les patients non porteurs d'une mutation pathogène, en une levée d'une angoisse illégitime et d'une prise en charge médicale lourde et injustifiée, puisque ces individus ont alors un risque similaire de développer un cancer du sein ou de l'ovaire à celui observé dans la population générale. Pour les patients porteurs d'une altération génétique constitutionnelle en lien avec le syndrome seins-ovaires, ceux-ci doivent se voir proposer une stratégie de suivi spécifique, basée sur la surveillance et/ou la chirurgie préventive et l'accès à de nouvelles thérapies ciblées dans un contexte de médecine personnalisée.

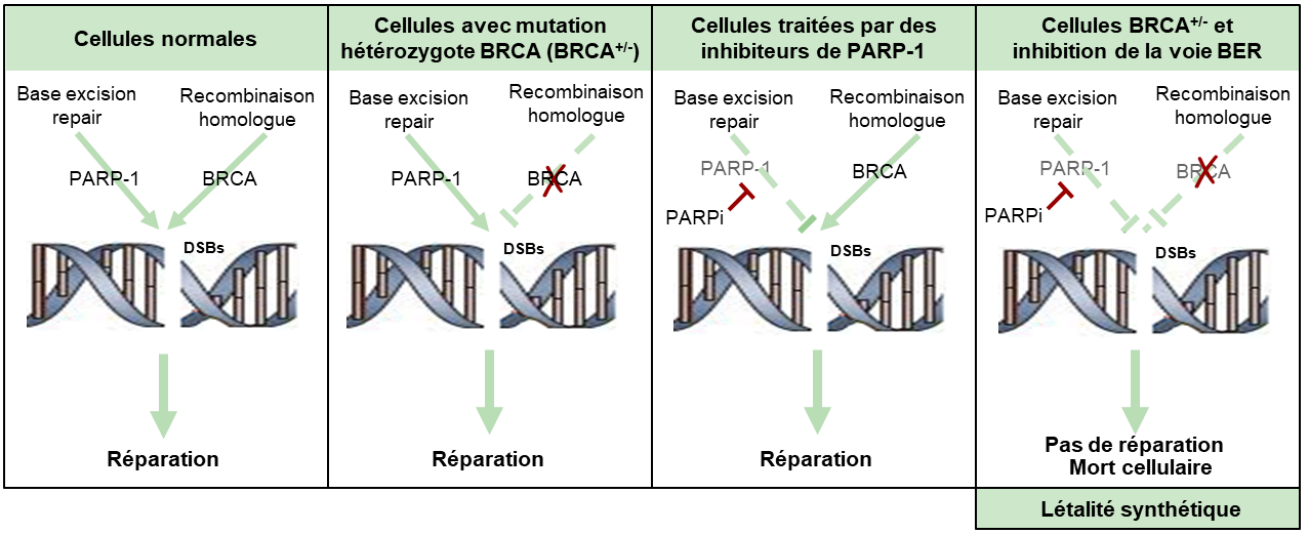
Les femmes à risque, c'est-à-dire les femmes porteuses d'une mutation *BRCA* ou *PALB2*, en particulier, peuvent se voir proposer une surveillance mammaire et gynécologique. Pour permettre un dépistage le plus précoce possible, ces femmes sont surveillées dès l'âge de 20 ans avec un suivi clinique biannuel puis sur un rythme annuel, par mammographie, par imagerie par résonnance magnétique (IRM) et éventuellement par échographie entre 30 et 65 ans, puis par mammographie +/- échographie uniquement au-delà de 65 ans et par des examens pelviens annuels ou semi-annuels à partir de 35 ans accompagnés d'un test sanguin pour détecter l'antigène CA125, biomarqueur du cancer de l'ovaire (pour revue : Kobayashi *et al.*, 2013). De plus, une chirurgie dite de réduction de risque peut être proposée à ces patientes. Il s'agit d'une mastectomie bilatérale et/ou une annexectomie prophylactiques (ou salpingo-oophorectomie bilatérale) qui permettent de réduire

très significativement le risque d'atteinte d'un cancer du sein ou de l'ovaire, réduction de 90% et 80% pour les cancers du sein et de l'ovaire, respectivement (pour revue : Kobayashi *et al.*, 2013).

De façon importante, l'identification de la mutation pathogène dans *BRCA1* ou *BRCA2* conditionne aujourd'hui l'accès à de nouvelles thérapies ciblées très prometteuses en particulier celles basées sur les inhibiteurs de PARPs (poly-(ADP-riboses) polymérases), eux-mêmes basés sur le concept de létalité synthétique (Figure 16 ; pour revue : McLornan *et al.*, 2014). Les PARP sont des enzymes impliquées dans la réparation des cassures simples brins l'ADN (voie BER, *base excision repair*), notamment en signalant après reconnaissance, cet ADN monocaténaire au système enzymatique chargé de restaurer l'ADN bicaténaire (Durkacz *et al.*, 1980; pour revue : Sonnenblick *et al.*, 2015). La voie de réparation de l'ADN dépendante des PARP est complémentaire de la voie de réparation par RH dans laquelle sont impliquées les protéines BRCA. Lorsque l'inhibiteur de l'enzyme PARP1 est administré à ces patientes, la voie de réparation dépendante des PARP est inactivée. Les cassures simple-brin non réparées du fait de l'inhibition de PARP1 se transforment alors en cassures double-brin au cours de la réplication de l'ADN. Dans les cellules tumorales, déficientes en BRCA, ces cassures doubles brin qui s'accumulent ne peuvent être réparées, ce qui entraine un arrêt du cycle cellulaire en G2/M, conduisant à l'apoptose des cellules tumorales déficientes en BRCA spécifiquement. Les cellules dites normales, qui elles possèdent une recombinaison homologue toujours fonctionnelle, peuvent procéder à la réparation des cassures doubles brins, restant viables. Récemment, il a d'ailleurs été montré que l'efficacité des traitements anti-PARP pourrait être accrue *via* l'utilisation d'un autre inhibiteur, l'anti RAD52, intervenant également dans la RH. On parle alors de double létalité synthétique (Sullivan-Reed *et al.*, 2018). A l'heure actuelle, les patientes ayant développé un cancer de l'ovaire associé à une variation pathogène des gènes BRCA sont éligibles pour un traitement d'entretien aux inhibiteurs de PARP. Les dernières études portant sur le premier d'entre eux, le lynparza™ (Olaparib), disposant d'une autorisation de mise sur le marché européenne depuis 2014, ont montré une amélioration de la survie sans progression de 2 ans par rapport au placébo (étude de phase III SOLO-2). De plus, l'intérêt des inhibiteurs PARP dans le traitement d'entretien des cancers du sein avec une variation pathogène BRCA est en cours d'essai par une étude clinique de phase III (étude OlympiAD). Aux États-Unis, deux autres inhibiteurs de PARP, le Zejula (Niraparib), qui a montré une réduction du risque relatif de progression de la maladie ou de décès de 74% chez les patientes porteuses de mutations du gène BRCA contre 55% chez les autres patientes en essai clinique de

phase III, et le Rubraca (Rucaparib) sont homologués en traitement de maintenance pour certains cancers de l'ovaire associés à une mutation BRCA, notamment des récidives du cancer épithélial de l'ovaire, du cancer des trompes de Fallope ou du carcinome péritonéal primaire chez les patientes encore sensibles à la chimiothérapie aux sels de platine et pour le traitement du cancer de l'ovaire évolué qui ont été traités avec deux ou plus de deux chimiothérapies, respectivement (Haute autorité de santé).

<u>Figure 16</u> : Mécanisme de la mort cellulaire spécifique des cellules déficientes en activité BRCA provoquée par la létalité synthétique, elle-même induite par l'inhibition de la poly(ADP-ribose) polymérase-1 (PARP-1) impliquées dans la voie de réparation par excision de base (adapté de Iglehart and Silver, 2009).

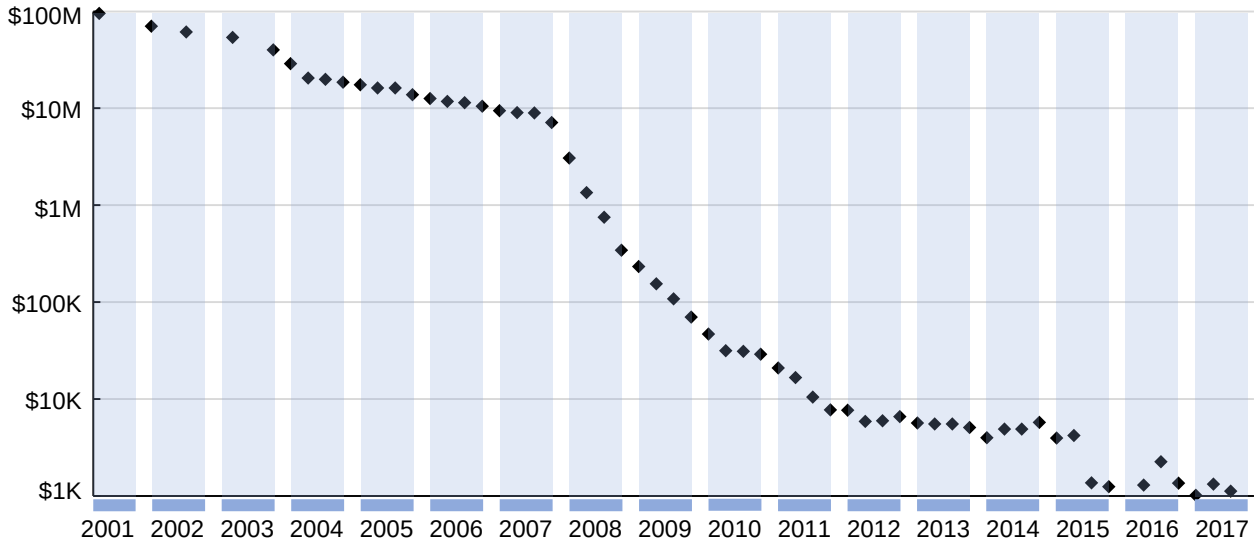


CHAPITRE II : PROBLEMATIQUE DE L'INTERPRETATION BIOLOGIQUE DES VARIATIONS

1) Implémentation du séquençage à haut-débit en génétique clinique

La notion de séquençage nouvelle génération (NGS, *next-generation sequencing*), est apparue au début des années 2000 avec l'arrivée sur le marché de la première génération de séquenceurs haut-débit, depuis remplacée par les deuxième et maintenant troisième générations, se renouvelant sans cesse pour séquencer toujours plus vite et toujours moins cher (Figure 17). Avant cela, il avait fallu 13 ans et 2,7 milliards de dollars entre 1990 et 2003 à un consortium (*Human Genome Project*) réunissant les seize plus grands instituts de biotechnologie de la planète pour parvenir à déchiffrer les 3,4 milliards de bases du génome humain (International Human Genome Sequencing Consortium, 2004; Lander *et al.*, 2001; Venter *et al.*, 2001). 15 années plus tard, cette même analyse peut être réalisée en vingt-quatre heures seulement pour 1000 dollars à peine (Figure 17 ; Mardis, 2011; Service, 2006).

Figure 17 : Evolution des coûts de séquençage d'un génome humain en fonction du temps. Données fournies par le NHGRI (*National Human Genome Research Institute*).



Aujourd'hui, la génétique humaine est entrée dans une ère nouvelle grâce au développement de nouvelles techniques de séquençage de l'ADN à haut débit (pour revue : Shendure, 2011). En effet, cette révolution technologique a considérablement facilité le criblage de mutations en diagnostic moléculaire mais a également permis de révéler l'existence d'une très grande variabilité interindividuelle du génome humain, notamment avec la détection d'un grand nombre de variations génétiques, structurelles (*copy number variation*, CNV) ou de séquence (SNV, *single nucleotide variant*), rares voire uniques, entre individus (1000 Genomes Project Consortium *et al.*, 2010; Lek *et al.*, 2016; gnomAD, http:// gnomad.broadinstitute.org/). Il est estimé, à partir du séquençage de 123 136 exomes, qu'environ 23 000 variations nucléotidiques sont détectées lors du séquençage de l'exome d'un individu, la plupart (environ 99% des SNV détectées) correspondant à des polymorphismes, variations avec une fréquence allélique élevée (supérieure à 1%) dans la population générale. Après élimination des variations polymorphiques répertoriées dans les bases de données de population contrôle, il demeure environ 1600 variations rares (dont la fréquence est inférieure à 0.1%) voire privées (spécifiquement et uniquement retrouvées dans le génome d'un individu) (*Genome Aggregation Database*, Mars 2017). Et chacune de ces variations est ainsi susceptible d'avoir un impact fonctionnel, pouvant alors être potentiellement pathogène.

Dans ce contexte de génétique médicale, l'enjeu n'est donc plus de détecter l'ensemble des variations présentes dans le génome d'un patient, mais plutôt d'identifier, parmi ces milliers de variations, celles potentiellement à l'origine de la maladie (pour revues : Cooper and Shendure, 2011; Frebourg, 2014). Cette problématique est particulièrement importante en oncogénétique et, notamment, dans le contexte des cancers héréditaires les plus fréquents, c'est-à-dire le syndrome seins-ovaires (gènes BRCA) et le syndrome de Lynch (gènes MMR).

2) Problématique des VSI dans les gènes MMR et BRCA

Le diagnostic du syndrome seins-ovaires et du syndrome de Lynch repose en partie sur un diagnostic moléculaire qui n'est réalisé qu'après identification d'un certain nombre de signes évocateurs de ces maladies. Il est axé sur la recherche, dans le génome d'un patient, d'une mutation constitutionnelle hétérozygote entrainant la perte de fonction d'un des gènes de prédisposition. L'identification de l'altération causale est essentielle pour le patient et ses apparentés du point de vue diagnostique, pronostique et thérapeutique. En effet, elle permet une optimisation de la prise

en charge et un suivi médical approprié du patient et des apparentés. Elle devrait également contribuer à éviter le phénomène de résistance à certains traitements et à conditionne aujourd'hui l'accès à de nouvelles thérapies ciblées très prometteuses éventuellement, basées, par exemple, sur le concept de létalité synthétique.

Cependant, une fraction très importante (~75-80%) de cas évocateurs de syndrome de Lynch et de syndrome seins-ovaires reste encore aujourd'hui sans explication moléculaire. Il est possible qu'une partie de ces cas inexpliqués soit due à des variations nucléotidiques identifiées chez les patients lors du diagnostic moléculaire mais actuellement classées comme des VSI (variations de signification inconnue). En effet, bien que les réarrangements génomiques dans les gènes MMR et les gènes BRCA expliquent 5-20% des cas de syndrome de Lynch et 9-18% du syndrome seins-ovaires, respectivement, la majorité des variations identifiées dans ces gènes correspond à des altérations nucléotidiques ponctuelles de type substitutions ou des insertions/délétions (indels) de petites tailles (pour revues : Kobayashi *et al.*, 2013; Peltomäki, 2014). Si le caractère délétère des mutations tronquantes (mutations non-sens et mutations à l'origine d'un décalage du cadre de lecture telles que les indels et les introniques localisées sur les positions les plus conservées des sites consensus d'épissage notamment) est le plus souvent évident, il est parfois difficile de statuer sur le caractère neutre ou délétère d'autres types de variations, dans le cas de ces syndromes en raison du large spectre mutationnel des gènes MMR et BRCA, de la pénétrance incomplète de certaines mutations, du manque de données de co-ségrégation et de l'absence de données fonctionnelles (Figure 18). Ces variations, dont l'interprétation clinico-biologique est beaucoup plus délicate, représentent environ 30% des variations détectées dans les gènes MMR et BRCA et constituent un obstacle majeur à l'optimisation de la prise en charge des patients et de leurs apparentés. Il s'agit principalement de variations de type faux-sens, d'insertions/délétions en phase, de variations silencieuses au niveau traductionnel ou encore de variations localisées dans les régions introniques, en dehors des positions les plus conservées des sites consensus d'épissage (-1, -2 et +1, +2). Les VSI ne pouvant pas être utilisées pour le conseil génétique, celles-ci représentent un obstacle majeur à l'optimisation de la prise en charge des patients et de leurs apparentés.

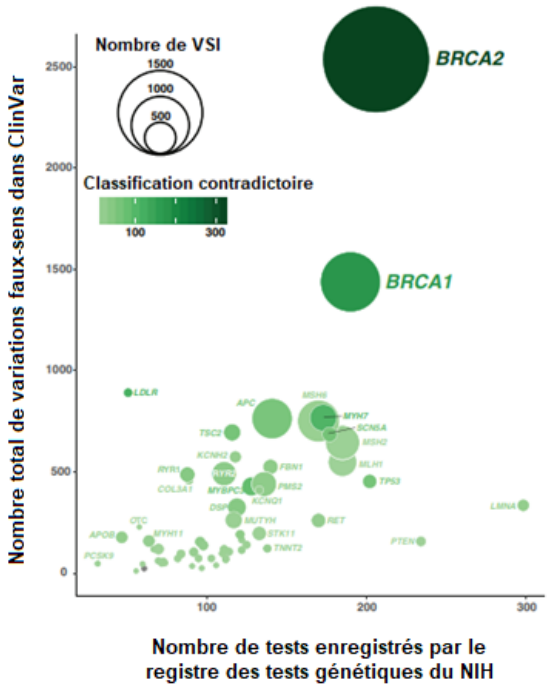


Figure 18 : Problématique des VSI en génétique médicale (d'après Starita *et al.*, 2017)**.** Beaucoup de variations faux-sens ont été découvertes et la plupart sont actuellement considérés comme des VSI (variations de signification inconnue). Le nombre de tests génétiques enregistrés corrèle avec le nombre de variations faux-sens identifiées (Spearman R^2 = 0.61). Les données intégrées dans ce graphique ont été extraites de la base de données ClinVar à la date du 5 Avril 2017.

3) Efforts nationaux et internationaux pour la classification des variations

C'est pourquoi de grands efforts nationaux et internationaux ont été déployés dans le but de collecter et classer les variations identifiées dans le génome des patients, conduisant à la création de groupes de travail et/ou de consortia et/ou de bases de données spécifiques. En France, l'ensemble des 25 laboratoires du dispositif national d'oncogénétique français, en charge de la réalisation des tests diagnostic des prédispositions génétiques à certaines formes de cancers, est regroupé autour du groupe de travail GGC (Groupe Génétique et Cancer), dont le développement est inscrit depuis 2003 au sein des trois plans cancer successifs par l'INCa (Institut National du Cancer). Sa mission est d'améliorer les connaissances sur les prédispositions héréditaires au cancer afin de garantir les bonnes pratiques de prise en charge des patients et de leur famille. Pour cela, des bases de données spécifiques et disponibles dans le domaine publique ont été développées afin de non seulement répertorier les variations identifiées dans les gènes MMR (UMD) et les gènes BRCA (BRCAShare) par l'ensemble des laboratoires français réalisant le diagnostic moléculaire du syndrome de Lynch et du syndrome seins-ovaires, mais également afin de collecter l'ensemble des informations génétiques (données de co-occurrence, de ségrégation et de fréquence), fonctionnelles (prédictions *in silico* et tests fonctionnels sur ARN et protéine), cliniques (phénotype) et tumorales (statut MMR ou ER/PR/HER2 tumoral) (Caputo *et al.*, 2012; Grandval

et al., 2013). L'ensemble de ces données contribuent à classer de manière homogène les variations selon 5 classes : classe 1 (variations neutres), classe 2 (variations probablement neutres), classe 3 (VSI), classe 4 (variations probablement causales) et classe 5 (variations causales).

Au niveau international, l'organisation scientifique internationale et multidisciplinaire InSiGHT (*International Society for Gastrointestinal Hereditary Tumours*), fondée en 2005, regroupe un comité de 45 experts de tous horizons travaillant sur la prédisposition aux cancers héréditaires gastro-instestinaux, notamment le syndrome de Lynch (Plazzer *et al.*, 2013). Depuis sa création, ces experts travaillent de concert pour standardiser la classification des variations identifiées notamment dans les gènes MMR en établissant des critères ou recommandations de classification s'appuyant essentiellement sur la nature de la variation, sa fréquence, les données fonctionnelles et familiales, et une analyse quantitative multifactorielle basée sur des données tumorales et des données de ségrégation et s'articulant, selon les recommandations de l'IARC (*International Agency for Research on Cancer*), autour d'un système a 5 classes de pathogénicité, sensiblement similaires à celles établies par le GGC : classe 1 (non pathogènes), classe 2 (probablement non pathogènes), classe 3 (VSI), classe 4 (probablement pathogènes) et classe 5 (pathogènes). Sur ces bases, le consortium international ENIGMA (*Evidence-based Network for the Interpretation of Germline Mutant Alleles*), a mis en place, à son tour en 2015, des recommandations standardisées de classification en 5 classes des variations identifiées dans les gènes de prédisposition au syndrome seins-ovaires notamment les gènes BRCA, en s'appuyant directement sur les recommandations définies par InSiGHT (Spurdle *et al.*, 2012). Il est à noter qu'auparavant l'IARC avait déjà proposé en 2008 un système de classification standardisé à 5 classes appliqué aux variations identifiées dans les gènes de prédisposition au cancer (Moghadasi *et al.*, 2016; Plon *et al.*, 2008).

4) Critères de classification des variations

La mise en place du NGS à visée diagnostique a confronté les laboratoires de diagnostic génétique à une augmentation des problématiques de l'interprétation des variations de séquence, au-delà des gènes MMR et BRCA. Cependant, en l'absence de recommandations, des nombreuses variations ont été rapportées causales ou bégnines sans arguments suffisants pour supporter cette classification, en particulier des variations identifiées dans des gènes dont les laboratoires respectifs

n'avaient pas ou peu d'expertise spécifique. En 2015, l'ACMG (*American College of Medical Genetics*) associée à l'AMP (*Association for Molecular Pathology*) ont établi, sur le plan international, des recommandations spécifiques pour la classification des variations identifiées dans les maladies Mendéliennes afin d'homogénéiser les pratiques des laboratoires de diagnostic (Amendola *et al.*, 2016; Richards *et al.*, 2015).

La classification clinique d'une variation repose sur un faisceau d'arguments dont le poids est plus ou moins important, en fonction de la nature de l'argument, dans l'interprétation du variant tels que PVS/PS/PM/PP désignent des arguments très fort (*pathogenic very strong*)/fort (*pathogenic strong*)/moyen (*pathogenic moderate*)/faible (*pathogenic supporting*) en faveur de la pathogénicité du variant et BA/BS/BP désignent les arguments individuel suffisant (*stand-alone*)/forts (*benign strong*)/faibles (*benign supporting*) en faveur du caractère bénin du variant (Richards *et al.*, 2015). Chacun des arguments est ainsi pondéré puis l'ensemble est combiné afin d'établir une probabilité de certitude qu'une variation soit pathogène ou bénigne, probabilité traduite en 5 classes de pathogénicité : classe 1 (variation bénigne, probabilité de neutralité >99%), classe 2 (variation probablement bégnine, probabilité de neutralité >90%), classe 3 (variation de signification inconnue), classe 4 (variation probablement pathogène, probabilité de pathogénicité >90%), classe 5 (variation pathogène, probabilité de pathogénicité >99%) (Richards *et al.*, 2015). Ce système de classification est similaire à celui établi par l'IARC et basé sur un système bayésien. Ce dernier consiste à générer une probabilité postérieure (PP, *posterior probability*) sur la pathogénicité de la variation analysée, probabilité également traduite en 5 classes : classe 1 (variation non pathogène ou sans signification clinique $PP \leq 0.1\%$), classe 2 (variation probablement non pathogène, de faible signification clinique $0.1\% < PP \leq 5\%$), classe 3 (VSI, $5\% < PP < 95\%$), classe 4 (variation probablement pathogène, $95\% \leq PP < 99\%$), classe 5 (variation pathogène, $PP \geq 99\%$) (Goldgar *et al.*, 2004; Plon *et al.*, 2008). Chacune de ces classes est alors associée à des recommandations spécifiques pour la prise en charge des patients et de leurs apparentés.

L'ensemble de ces systèmes de classification des variations s'appuient aujourd'hui sur un ensemble d'arguments (génétiques, familiaux, fonctionnels, *in silico*, cliniques ou tumoraux) rarement suffisants à eux même pour statuer sur la pathogénicité du variant, mais qui, pris dans leur globalité (effet additif) permettent d'interpréter une variation. Ces arguments (ainsi que leur contribution sont détaillés dans Richards *et al.*, 2015 et Tableau 5) :

Tableau 5 : Recommandations de l'ACMG-AMP pour la classification des variations (d'après Jarvik and Browning, 2016)

	Bénin		Pathogène			
	Suffisant/Fort	Faible	Faible	Moyen	Fort	Très fort
Données épidémiologiques	Fréquence allélique trop importante par rapport à la fréquence de la pathologie (**BA1/BS1**) Présence de la variation chez les contrôles incohérente avec la pénétrance de la pathologie (**BS2**)			Variation absente les bases de données de populations contrôles (**PM2**)	Prévalence de la variation chez les individus atteints significativement supérieure à celle des contrôles (**PS4**)	
Données structurales		Variation faux-sens prédite sans effet par l'ensemble des logiciels de prédictions de pathogénicité interrogés (**PB4**) Variation faux-sens dans un gène où seules les variations tronquantes sont associées à la pathologie (**BP1**) Variations synonymes sans impact prédit sur l'épissage, (**BP7**) Indels en phase dans une région répétée sans fonction connue (**BP3**)	Variation faux-sens prédite délétère par l'ensemble des logiciels de prédiction de pathogénicité interrogés (**PP3**)	Variation à l'origine d'un changement d'acide aminé différent à la même position qu'une variation faux-sens pathogène connue (**PM5**) Variation affectant la longueur de la protéine (**PM4**)	Variation à l'origine du même changement d'acide aminé qu'un variant pathogène connu (**PS1**)	Variation ayant un effet nul prédit dans un gène où la perte de fonction est un mécanisme pathogène connu (**PVS1**)
Données fonctionnelles	Etudes fonctionnelles bien établies montrant un impact non délétère de la variation (**BS3**)		Variation faux-sens dans un gène avec un faible taux de faux-sens bénins et dans lequel les faux-sens sont un mécanisme responsable de la pathologie fréquent (**PP2**)	Variation située sur un *hot spot* mutationnel ou un domaine fonctionnel essentiel exempt de variations bénignes (**PM1**)	Etudes fonctionnelles bien établies montrant un impact délétère de la variation (**PS3**)	
Données de ségrégation	Variation ne ségrégant pas avec la pathologie chez les apparentés (**BS4**)		N≤1/8 ou N≤1/4 si données issues d'une ou plusieurs familles (**PP**)	N≤1/16 ou N≤1/8 si données issues d'une ou plusieurs familles (**PM**)	N≤1/32 ou N≤1/16 si données issues d'une ou plusieurs familles (**PS**)	
Données *de novo*				Variation *de novo* sans confirmation de la paternité et de la maternité (**PM6**)	Variation *de novo* avec confirmation de la paternité et de la maternité (**PS2**)	
Données alléliques		Variation observée en cis ou en *trans* avec une autre variation pathogène, si la pathologie à pénétrance complète a une transmission autosomique dominante ou liée à l'X (**BP2**)		Variation observée en *trans* avec une autre variation pathogène, si la pathologie a une transmission récessive (**PM3**)		
Autres bases de données		Source documentée classant cette variation comme bénigne (**BP6**)	Source documentée classant cette variation comme pathogène (**PP5**)			
Données additionnelles		Variation en co-occurrence avec une variation pathogène dans un autre gène impliqué dans la pathologie (**BP5**)	Le phénotype du patient ou les antécédents familiaux sont très spécifiques pour le gène (**PP4**)			

a) aux données bibliographiques. Il est nécessaire, en premier lieu, de vérifier si la variation a déjà été reportée dans la littérature (PubMed) et/ou dans des bases de données de variations génétiques détectées chez des patients (ClinVar, HGMD, LOVD, COSMIC ou bases de données spécifiques de gènes ou de maladies).

b) aux données de population. Issues de bases de données spécifiques (1000 genomes, ExAC, gnomAD, ESP), ces données sont utilisées pour déterminer la fréquence allélique du variant mis en évidence dans la population générale, une fréquence allélique importante (supérieure à la prévalence de la maladie) étant en faveur du caractère bénin du variant.

c) aux prédictions *in silico*. Elles peuvent être générées à deux niveaux : (i) l'épissage de l'ARN avec des outils qui permettent d'appréhender l'effet potentiel d'une variation sur les sites d'épissage (par exemple, MaxEntScan, SpliceSiteFinder, NNSPLICE, GeneSplicer et Human Splicing Finder) et (ii) sur la protéine avec des algorithmes qui permettent d'appréhender la pathogénicité d'une variation faux sens en fonction de la conservation de la base nucléotidique et surtout de l'acide aminé dans l'évolution, de la distance physico-chimique entre le résidu aminoacide sauvage et muté et/ou de la position dans un domaine fonctionnel connu de la protéine (par exemple, SIFT, MutationTaster, Polyphen-2 ou A-GVGD).

d) aux données fonctionnelles. Des analyses fonctionnelles génériques (viabilité, sensibilité à des drogues, etc) ou spécifiques, sur l'ARN (tests fonctionnels d'épissage) ou sur la fonction de la protéine (essais fonctionnels spécifiques à la protéine), *in vivo* ou *in vitro* bien établies peuvent contribuer à démontrer un impact délétère ou neutre de la variation.

e) aux données de co-occurrence. Elles réfèrent à l'existence de variations associées à la variation d'intérêt. Si des génotypes homozygotes ou hétérozygotes composites pour des variations pathogènes sur certains gènes sont supposés être létaux au stage embryonnaire, alors des variations identifiées dans cette configuration peuvent être classées neutres sur la base de cette observation.

f) aux données de ségrégation. Il s'agit de déterminer si la présence de la variation d'intérêt chez les individus atteints (dans une même et/ou plusieurs familles), soit due au hasard plutôt que due à une co-ségrégation (Jarvik and Browning, 2016). Le poids de cet argument est d'autant plus important que le nombre de sujets testés attestant de la co-ségrégation du génotype avec le phénotype est élevé ou les sujets testés éloignés dans l'arbre généalogique.

A noter que chez un individu atteint mais sans antécédent familial, la variation peut être apparue *de novo* chez l'individu.

g) aux données cliniques ou généalogiques. Il s'agit de déterminer si le phénotype ou l'histoire familiale sont en faveur de la pathologie associée au gène portant la variation d'intérêt. La force de cet argument est fonction de la spécificité du phénotype observé, de l'hétérogénéité génétique pour ce phénotype et du nombre de gènes analysés.

h) aux données somatiques (argument absent dans les recommandations de l'ACMG). L'interprétation de variations constitutionnels identifiés chez des patients évocateurs de cancers héréditaires peut aussi s'appuyer sur des analyses somatiques quand du tissu tumoral est disponible.

Malgré ces recommandations, la pathogénicité d'une grande partie des variations reste actuellement évaluée sur la base de l'altération de la fonction ou de la structure de la protéine par des analyses expérimentales et/ou bioinformatiques uniquement. Or, le comité d'interprétation des variations d'InSiGHT a mentionné, dans l'organigramme d'interprétation fonctionnelle que l'analyse des transcrits doit précéder les analyses protéiques au moins pour les indels en phase et les variations faux-sens, et doit être réalisée pour toutes les variations introniques et les variations synonymes *a priori* sans impact sur la protéine (Thompson *et al.*, 2014). En effet, si certaines variations sont susceptibles d'affecter directement la protéine, toutes peuvent potentiellement modifier un processus en amont, l'épissage des ARN pré-messagers (pré-ARNm). Il a d'ailleurs été montré qu'une fraction importante (~35%) des variations qui altèrent l'épissage sont à l'origine de nombreuses maladies génétiques (Tableau 6 ; pour revues : Baralle *et al.*, 2009; Cooper *et al.*, 2009; Durand *et al.*, 2007; Wang and Cooper, 2007). Il est donc probable qu'une fraction importante des VSI, en particulier celles détectées dans le syndrome de Lynch et le syndrome seins-ovaires soient responsables d'une altération de l'épissage.

Tableau 6 : Fréquence des mutations d'épissage dans des maladies génétiques héréditaires (d'après Baralle *et al.*, 2009). Ces données ont été générées à partir de la base de données publique HGMD, *Human Gene Mutation Database* (24/11/2008).

Maladie/Phénotype	Gène	#Epissage / # total (%)
Ataxie-télangiectasie	*ATM*	18
Syndrome seins-ovaires	***BRCA1***	**9**
Medium chain acyl CoA dehydrogenase deficiency	*CADM*	10
Mucoviscidose	*CFTR*	14
Dystrophie musculaire de Duchenne	*DMD*	9
Maladies du sang (Thalassémies, anémies)	*HBA1/2*	3
Maladies du sang (Thalassémies, anémies)	*HBB*	10
Déficit en hypoxanthine-guanine phosphoryl transférase	*HPRT*	15
Dysautonomie familiale	*IKBKAP*	33
Parkinson et demenses frontotemporal	*MAPT*	33
Syndrome de Lynch	***MLH1***	**18**
Syndrome de Lynch	***MSH2***	**9**
Neurofibromatose de Type 1	*NF1*	19
Neurofibromatose de Type 2	*NF2*	22
Rétinite pigmentaire	*RHO*	3
Atrophie musculaire spinale	*SMN1/2*	4
Tumeurs de Wilms	*WT1*	11

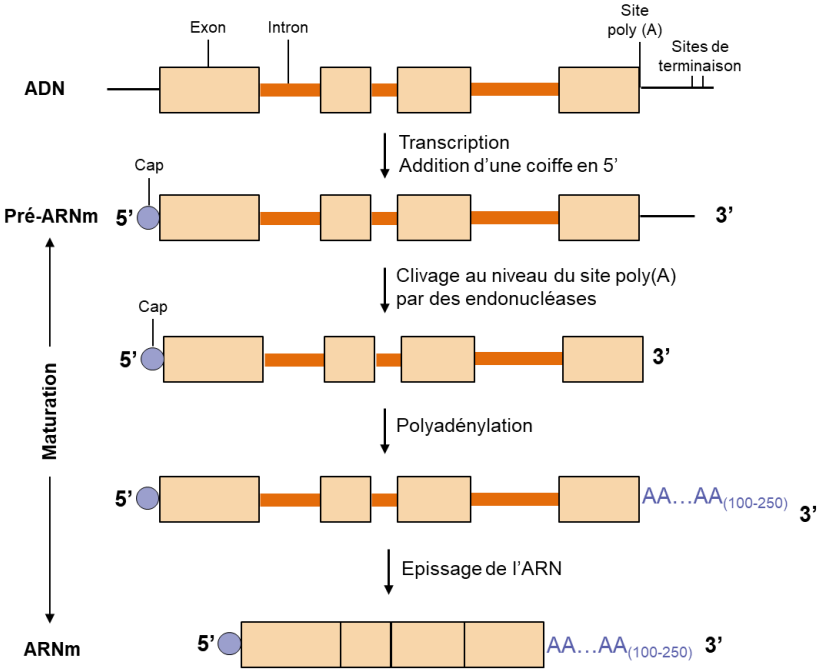
CHAPITRE III : MECANISME D'EPISSAGE DE L'ARN PRE-MESSAGER

1) Définition du processus d'épissage

En 1977, Richard Roberts et Phillip Sharp ont tous deux indépendamment, au même moment, avec les mêmes techniques et le même model (adénovirus), démontré pour la première fois la théorie du gène fragmenté : une partie du matériel génétique est retiré de l'ARN avant qu'il ne soit traduit en protéines (Berget *et al.*, 1977; Chow *et al.*, 1977). En effet, quand Roberts et Sharp comparèrent l'ARNm dans le noyau et dans le cytoplasme, ils le trouvèrent différent. En mettant en parallèle des segments d'ADN avec les segments d'ARN cytoplasmiques correspondant – en formant des hybrides ARN-ADN – ils trouvèrent qu'ils ne s'assemblaient pas complètement. Certaines régions correspondaient alors que d'autres semblaient avoir complètement disparues dans l'ARN cytoplasmique (Berget *et al.*, 1977; Chow *et al.*, 1977). L'ensemble de ces travaux ont été récompensés par un prix Nobel de Physiologie ou Médecine en 1993 (Carr, 1993).

Cette découverte a ensuite contribué à une évolution importante de la définition des gènes composé, selon Walter Gilbert, de régions intragéniques ou introns séparant des séquences exprimées nommées exons (Gilbert, 1978a; Sharp, 2005). Ces notions ont par la suite conduit à la découverte d'un nouveau processus biologique, le processus d'épissage. Avec la synthèse de la coiffe (*capping*) en 5' et la polyadénylation de l'ARNm en 3', l'épissage constitue une des étapes essentielles dans la maturation des ARN précurseurs des cellules eucaryotes (Figure 19). Ce processus consiste en l'élimination précise des introns des pré-ARNm et la liaison ou assemblage efficace des exons entre eux (pour revue : Cartegni *et al.*, 2002). Une fois correctement maturé dans le noyau, l'ARNm est ensuite exporté dans le cytoplasme où il sera traduit en protéines. D'ailleurs, il a été récemment montré qu'environ 80% des réactions d'épissage se déroulent de façon co-transcriptionnelle dans les cellules humaines (Girard *et al.*, 2012).

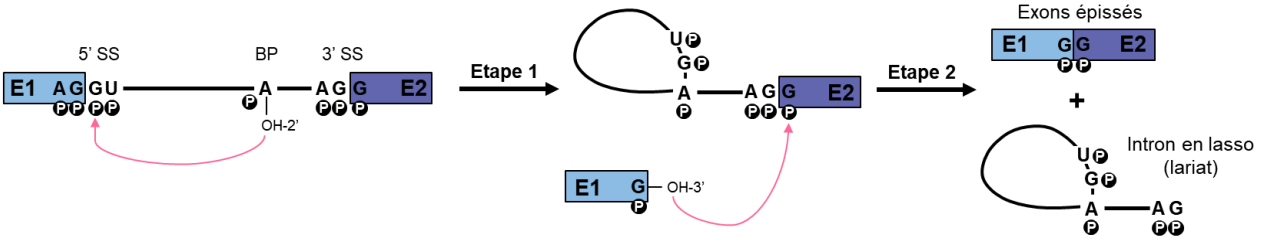
Figure 19 : Maturation des ARNs pré-messagers eucaryotes



2) La réaction catalytique d'épissage

Le mécanisme d'épissage comprend deux réactions de trans-estérification consécutives conduisant à la ligation des deux exons entre eux et à la libération de l'intron sous la forme d'une structure en lasso (*lariat*) (Figure 20). La première réaction de trans-estérification conduit à la formation d'une liaison 2'-5' phosphodiester entre la guanine du site 5' d'épissage et l'adénine du point de branchement grâce à une attaque nucléophile du groupement 2'-OH du ribose de cette adénosine vers le phosphate de la jonction exon-intron en amont. Il y a alors formation du lariat et la libération de l'exon en amont. La deuxième réaction de trans-estérification résulte en la formation d'une liaison phosphodiester entre les deux exons grâce au groupement 3'OH libre de l'exon en amont qui attaque le phosphate de la jonction intron-exon en aval (site 3' d'épissage en aval). Cela permet la liaison des exons entre eux et ainsi la production d'un ARNm mature (constitué d'exons uniquement), et la libération des introns sous forme de lasso qui sera par la suite ouvert par l'enzyme de débranchement 1 (DBR1) avant d'être dégradé par des ribonucléases (pour revues : De Conti *et al.*, 2013; Will and Lührmann, 2011).

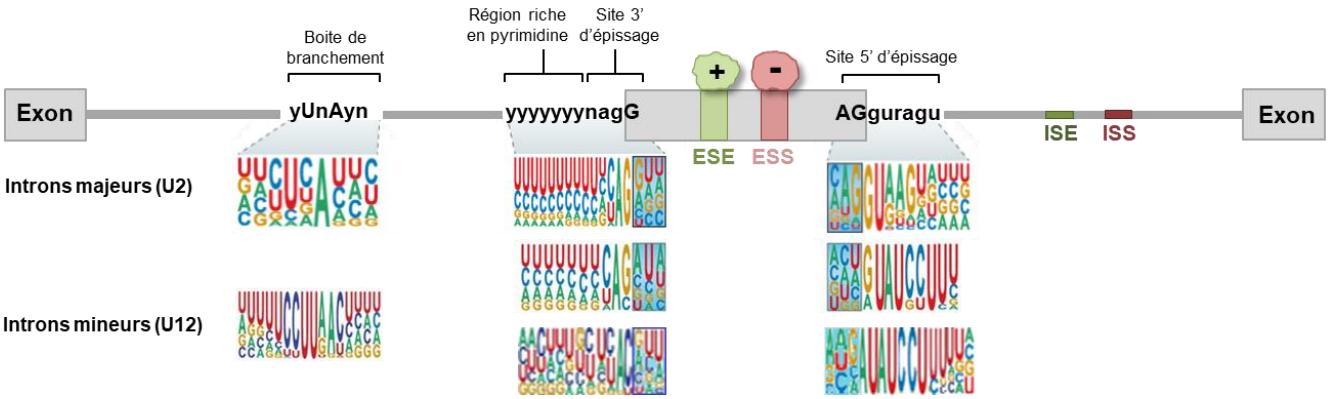
Figure 20 : Les réactions de transestérification dans l'épissage du pré-ARNm (adapté de Will and Lührmann, 2011). Les étapes 1 et 2 correspondent, respectivement, aux première et deuxième réactions de trans-estérification. SS, *splice site* ; BP, *branch point*, E, exon.



3) Les principaux signaux d'épissage

Plusieurs signaux spécifiques présents le long de la séquence des pré-ARNm contribuent à la définition des introns et des exons (Figure 21). Ces séquences incluent : (i) les sites 5' donneurs et 3' accepteurs d'épissage (5'ss, *5' splice site* ; 3'ss, *3' splice site*) situés en 5' et 3' de l'intron, à la jonction exon/intron ou intron/exon, respectivement, et définis par la présence de séquences consensus courtes, (ii) le point de branchement (BP, *branch point*) situé en amont du site 3' accepteur et (iii) la région riche en pyrimidines (PPT, *polypyridine tract*) située entre le point de branchement et le site 3' d'épissage (pour revue : Cartegni *et al.*, 2002). L'ensemble de ces signaux sont nécessaires et indispensables pour la reconnaissance, par la machinerie d'épissage, des séquences exoniques à épisser (pour revues : Wahl *et al.*, 2009; Will and Lührmann, 2011).

Figure 21 : Représentation schématique des signaux d'épissage. ESE, *exon splicing enhancers* ; ESS, *exon splicing silencers* ; ISE, *intron splicing enhancers* ; ISS, *intron splicing silencers*.



Les caractéristiques de ces signaux varient en fonction de la nature des séquences introniques. En effet, il existe, chez la majorité des eucaryotes, 2 types d'introns : les introns dits majeurs ou U2, présents chez tous les eucaryotes et les introns dits mineurs ou U12 (Jackson, 1991; Jenkins *et al.*, 1990; pour revue : Padgett, 2015). Dans les introns majeurs, représentant plus de 99,5% des introns humains, la grande majorité des sites 5' (donneurs) et 3' (accepteurs) d'épissage se caractérise, respectivement, par les séquences de référence ou séquences consensus dégénérées CAG|GURAGU et YAG|G qui s'étendent de la position -3 à la position +6 et de la position -3 à la position +1 (les séquences soulignées correspondent aux dinucléotides introniques les plus conservés, R pour purine et Y pour pyrimidine) (Figure 21 ; pour revues : Cartegni *et al.*, 2002; Fredericks *et al.*, 2015). Le site 3' d'épissage se caractérise plus précisément par l'existence de 3 éléments présents dans les 40 nucléotides bordant la jonction intron/exon : le site d'épissage 3' en lui-même, le PPT et la boite de branchement contenant le BP. La boite de branchement se situe généralement entre 18-40 nucléotides du site 3' d'épissage, pour 90% des boites de branchement identifiées (Gao *et al.*, 2008; Mercer *et al.*, 2015). Elle se définie selon la séquence consensus yUnAy, initialement identifiée par la comparaison des séquences des lariats obtenus pour 52 introns issus d'une vingtaine de gènes de référence, puis par cartographie fonctionnelle sur l'ensemble du génome (Figure 21 ; Gao *et al.*, 2008; Mercer *et al.*, 2015). Plus particulièrement, la boite de branchement contient le site de branchement proprement dit, le nucléotide indispensable à la réalisation de la première réaction de trans-estérification de la réaction d'épissage. Ce nucléotide correspond dans 92% des sites de branchement identifiés à une adénosine. La région riche en pyrimidine contient, quant à elle, une grande densité de pyrimidines et se situe jusqu'à 10-12 nucléotides en amont de la jonction intron-exon (Gao *et al.*, 2008; Mercer *et al.*, 2015).

Les dinucléotides introniques GU et AG des sites 5' et 3' d'épissage étant presque invariables dans les introns majeurs, ces derniers sont souvent appelés les introns GU-AG. Les dinucléotides introniques GU-AG caractérisent 99% des introns (pour revues : Fredericks *et al.*, 2015; Sibley *et al.*, 2016). Il existe cependant des sites d'épissage de type U2 dits atypiques (pour revues : Fredericks *et al.*, 2015; Sibley *et al.*, 2016). Les plus fréquents sont les sites d'épissage GC-AG, qui représentent environ 0,9% des sites d'épissage humains (Sheth *et al.*, 2006; Thanaraj and Clark, 2001; pour revue : Sibley *et al.*, 2016). Ces derniers, plus fréquents dans les introns alternatifs comparativement aux introns constitutifs, sont intrinsèquement plus faibles que les sites d'épissage GT parce que la substitution T>C à la position +2 de l'intron induit un mésappariement avec le

snRNP U1, même si le reste du motif des sites d'épissage GC tend à compenser le mésappariement (Thanaraj and Clark, 2001). De ce fait, les dinucléotides GT à la jonction exon-intron peuvent parfois être remplacés par les dinuclétotides GC sans altérer la précision de l'épissage mais en altérant sa vitesse (Aebi *et al.*, 1987).

Les introns mineurs ou introns U12, dont la fréquence d'occurrence chez les vertébrés reste faible, 0.15-0.34% (~700 introns), comparativement aux introns U2, se caractérisent également par la présence des sites 5' et 3' consensus d'épissage ainsi que la boîte de branchement (Figure 21 ; Patel and Steitz, 2003a; Sibley *et al.*, 2016). Mais à la différence des introns majeurs, ils ne possèdent pas de région riche en pyrimidines proprement dite (pour revues : Patel and Steitz, 2003b; Turunen *et al.*, 2013; Will and Lührmann, 2005) Les sites d'épissage U12 ont d'abord été définis par une combinaison inhabituelle mais très conservée de dinucléotides non canoniques AU et AC aux extrémités 5' et 3', respectivement. Pour cette raison, ils étaient à l'origine appelés introns ATAC. Plus tard, les introns U12 ont été décrits comme comprenant à la fois les dinucléotides AU-AC et GU-AG. En fait, la majorité des introns U12, environ 70 %, correspondent à des introns GU-AG chez l'homme. Il est à noter que d'autres combinaisons de dinucléotides ont été rapportées, en particulier au niveau du site 3' d'épissage AT-AC, AT-AA, AT-AG, AT-AT, parce que le site accepteur est plus tolérant aux substitutions. La boite de branchement des introns U12 est particulièrement conservée, contrairement à celle des introns U2, et se situe à 11-13 nucléotides du site 3' d'épissage. Elle joue un rôle déterminant dans l'identification du site 3' d'épissage U12 par la machinerie d'épissage.

Ces introns, découverts assez récemment, dans les années 1990, présentent une taille moyenne similaire aux introns majeurs chez l'homme (~3000-4000 nucléotides de longueur), bien qu'il existe des petits introns d'environ 90 nucléotides et moins, qu'on ne retrouve pas dans les introns mineurs (pour revues : Patel and Steitz, 2003b; Turunen *et al.*, 2013; Will and Lührmann, 2005). Les introns de type U12 diffèrent en particulier par les caractéristiques des signaux d'épissage et sont restreints à des gènes assurant des fonctions cellulaires essentielles, en particulier la réplication et la réparation de l'ADN, la transcription, le *processing* de l'ARN et la traduction. Dans ces gènes, seulement un intron de type U12 est généralement présent et coexiste avec des introns de type U2, même si une cinquantaine de gènes en possèdent deux et, une exception, le gène *NHE-6*, en possède trois. Il a été suggéré que ces introns pourraient réguler l'expression de

ces gènes. En effet, il a été montré que l'épissage co-transcriptionnel des introns U12 apparait comme étant significativement plus lent, au moins deux fois plus lent, comparativement à celui des introns U2.

4) Le splicéosome, principal acteur du processus d'épissage

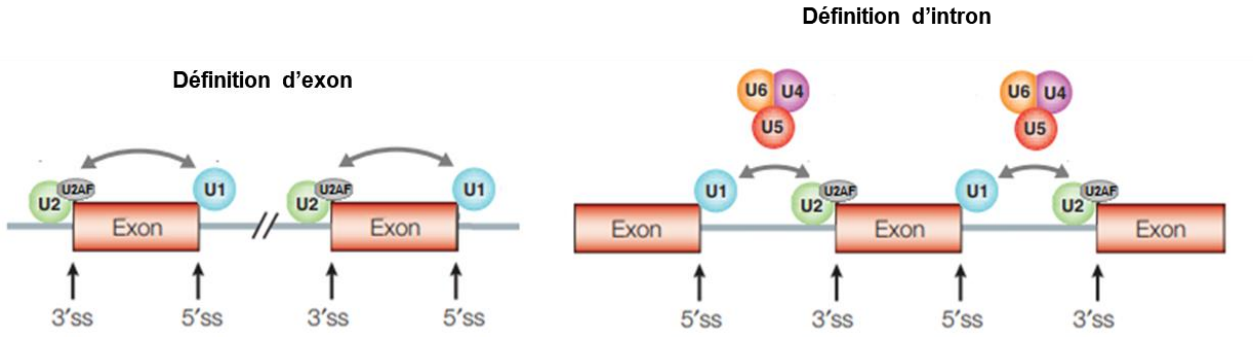
L'épissage de l'ensemble des introns est assuré par un complexe nucléaire ribonucléoprotéique (RNP, *ribonuclear proteins*) enzymatique très complexe, le splicéosome. Ce dernier correspond une macromolécule de plusieurs mégadaltons, composé de cinq petites ribonucléoprotéines nucléaires (snRNPs, *small nuclear ribonucleoproteins*) et d'un large nombre (entre 150 et 300) de facteurs protéiques différents non-snRNPs (Akerman *et al.*, 2015; pour revue : Papasaikas and Valcárcel, 2016; Will and Lührmann, 2011). Les snRNPs correspondent à de petites molécules d'ARN non codant (snRNA, *small nuclear RNA*) de 100 à 200 nucléotides de longueur associées à une dizaine de protéines différentes. Ces derniers forment des interactions ARN-ARN par appariement de base au niveau des séquences conservées des signaux d'épissage qui flanquent les introns et forment également des interactions entre eux par appariement de base au sein du splicéosome (pour revues : Nilsen, 1998; Padgett, 2005). De plus, les snRNPs apportent au splicéosome des facteurs protéiques spécifiques, avec lesquels ils sont associés via des interactions ARN-protéines. Ces facteurs, avec l'ensemble des protéines non-snRNPs, facilitent la liaison des snRNPs aux sites d'épissage et aussi la dynamique d'interactions ARN-ARN, ARN-protéine et protéine-protéine au niveau du splicéosome (pour revues : Chen and Cheng, 2012; Will and Lührmann, 2011). Il s'agit ainsi d'une des macromolécules les plus complexes dans les cellules eucaryotes (Akerman *et al.*, 2015; pour revue : Wahl *et al.*, 2009). La composition et la conformation du splicéosome sont très dynamiques et réversibles au court du temps, offrant à la machine d'épissage sa précision et sa flexibilité. En effet, le réseau ARN-protéines se forme et est constamment réarrangé durant l'assemblage du splicéosome et la catalyse avec un échange remarquable de protéines d'une étape à l'autre durant l'épissage, accompagné d'un remodelage des snRNPs (pour revue : Wahl *et al.*, 2009).

a. Le splicéosome majeur

Le splicéosome majeur, responsable de l'épissage des introns majeurs, soit environ 99% des introns contenus dans les pré-ARNm chez l'homme, se compose de 5 snRNPs : U1, U2, U4, U5 et U6 et de plusieurs facteurs non-snRNPs correspondant à des protéines de liaisons à l'ARN (RBPs, *RNA binding proteins*) notamment U2AF, SF1 et SRFs et des enzymes, en particulier des ARN hélicases, des kinases et des phosphatases (pour revues : Cooper *et al.*, 2009; Will and Lührmann, 2011). Il est important de noter qu'aucune des molécules snRNPs ne possède de centre catalytique. Ces molécules ne peuvent donc pas à elles seules assurer les réactions de trans-estérification. Les centres catalytiques se forment en réalité au fur et à mesure de la réaction d'épissage, en même temps que le splicéosome se forme, grâce aux interactions séquentielles des snRNAs et des protéines du splicéosome avec le pré-ARNm, leur substrat, et entre eux (pour revue : Will and Lührmann, 2011).

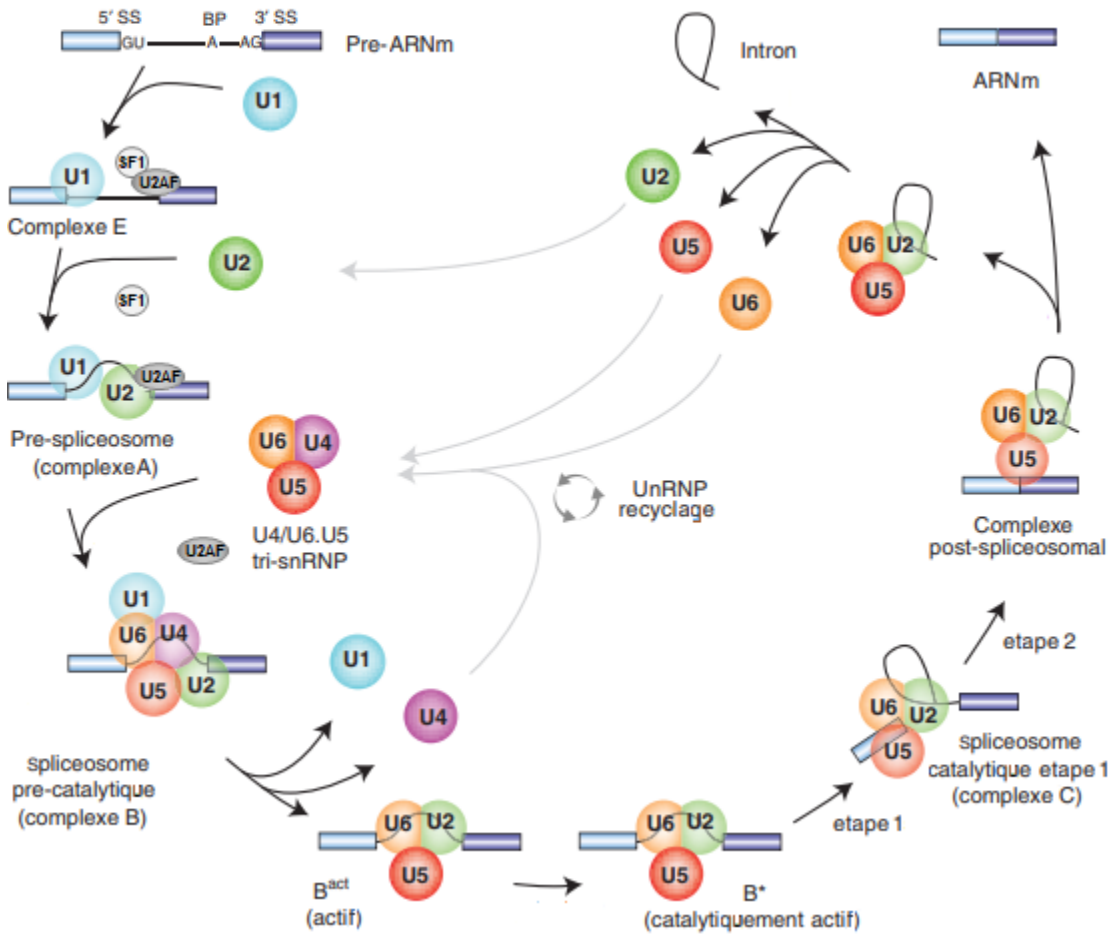
Chez les eucaryotes, deux modèles de reconnaissance des signaux d'épissage par le splicéosome majeur coexistent et sont fonction de la taille des introns : le modèle de définition d'intron ou d'exon (Figure 22). Le modèle de définition d'intron est responsable de l'épissage des introns dits de petite taille, introns dont la taille moyenne varie entre 200 et 250 nucléotides (Fox-Walsh *et al.*, 2005; pour revue : De Conti *et al.*, 2013). Pour ces introns, l'appariement entre les sites d'épissage a lieu via l'intron court qui sépare de longs exons. Le modèle de définition d'exon est, quant à lui, responsable de l'épissage des introns de grande taille, dont la taille dépasse les 250 nucléotides. Pour ces introns, la reconnaissance des signaux et l'assemblage de la machinerie d'épissage se fait au niveau de l'exon, où communiquent l'ensemble des facteurs d'épissage (Figure 22 ; Berget, 1995; Fox-Walsh *et al.*, 2005; pour revue : De Conti *et al.*, 2013).

Figure 22 : Complexes de définition d'exon et de définition d'intron (adapté de Daguenet *et al.*, 2015).



Dans le modèle de définition d'intron, le snRNP U1 se fixe au niveau du site 5' d'épissage et des facteurs non-snRNP, en particulier SF1/mBBP (*splicing factor 1/branch point binding protein*) et l'hétérodimère U2AF65/U2AF35 (U2AF, *U2 auxillary factor*), interagissent avec le point de branchement et la région riche en pyrimidine, respectivement, formant le complexe E (Figure 23; Fox-Walsh *et al.*, 2005; pour revue : De Conti *et al.*, 2013). Le snRNP U2 est alors recruté au niveau du point de branchement, formant le complexe A appelé aussi pré-splicéosome. La fixation du snRNP U2 au site de branchement nécessite la libération du facteur SF1/mBBP et est stabilisée par les composants protéiques de ce snRNP, notamment les facteurs SF3a et SF3b. Le tri-snRNP U4/U6.U5, qui est préassemblé à partir des snRNPs U5 et U4/U6 est ensuite recruté, générant le complexe pré-catalytique B. Pour que ce dernier devienne actif, des réarrangements majeurs dans les interactions snARN-ARN et ARN-protéine sont nécessaires. Ces réarrangements conduisent alors à la déstabilisation des snRNP U1 et U4, donnant naissance au splicéosome activé (aussi appelé complexe de Bact), puis au complexe B* après activation catalytique par l'hélicase Prp2 de la DEAH-box. Ce dernier peut alors catalyser la première réaction de trans-estérification, générant le complexe C qui, à son tour, catalyse la deuxième réaction de transestérification d'épissage. Le splicéosome se dissocie ensuite pour libérer l'ARNm mature et, après un remodelage supplémentaire, les snRNPs qui participent à d'autres réactions d'épissage et sont ainsi recyclées (pour revues : Wahl *et al.*, 2009; Will and Lührmann, 2011). Il est à noter que des ARN hélicases de la famille de DExD/H-box interviennent dans la majorité des étapes d'assemblage et d'activation du splicéosome afin de faciliter les changements de conformation de l'ARN et le remodelage des interactions ARN-ARN et ARN-protéines (pour revue : Will and Lührmann, 2011).

Figure 23 : Les étapes d'assemblage du splicéosome majeur dans le modèle de définition d'intron (adapté de Will and Lührmann, 2011).



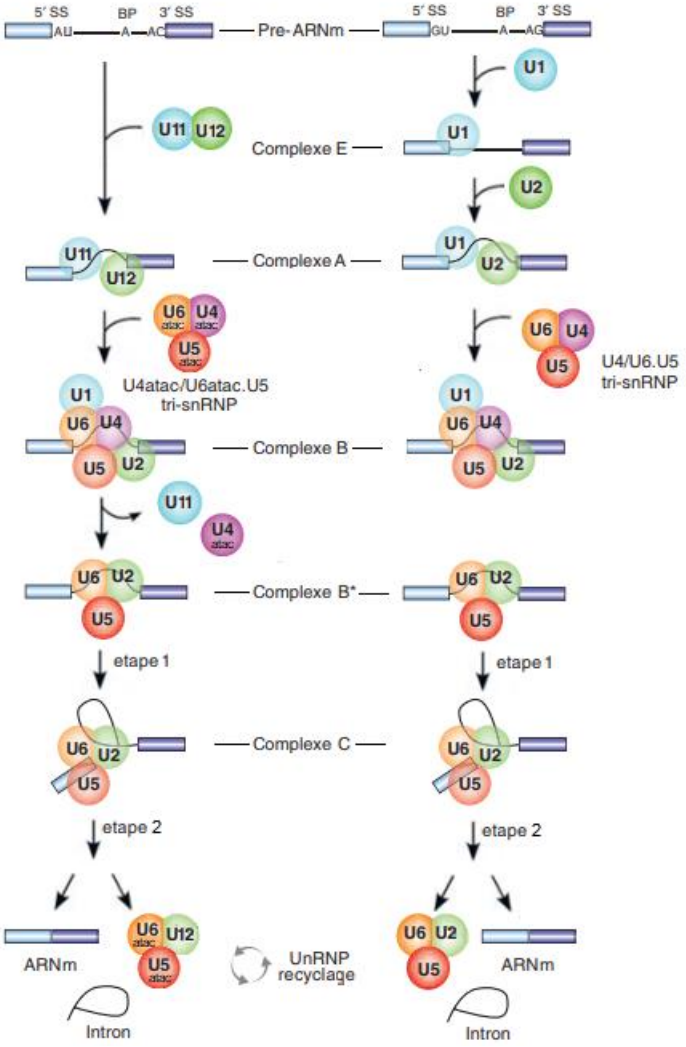
Dans la grande majorité des gènes humains, les introns étant de grande taille et séparant des petits exons, le modèle de définition d'exon est donc le plus répandu chez l'Homme. Dans ce modèle, les facteurs d'épissage snRNP U1 et U2AF se fixent également au niveau du site 5' d'épissage et de la région riche en pyrimidine, respectivement, mais cette interaction se déroule à travers l'exon (Figure 22, pour revue : Daguenet *et al.*, 2015). Intervient ensuite le snRNP U2, également recruté au niveau du point de branchement mais de manière concomitante au recrutement des protéines SR (*serine-arginine rich proteins*) au niveau des éléments exoniques activateurs d'épissage dit ESE (*exonic splicing enhancer*) (voir section III.5.1). Dans leur ensemble, ces protéines forment un réseau de stabilisation dans l'exon, qui soutient le complexe de définition d'exon. Après la reconnaissance des sites d'épissage, le complexe de définition d'exon

se convertit en modèle de définition d'intron (pour revue : Will and Lührmann, 2011). Il a été suggéré que le complexe de définition d'exon se convertit en complexe A, décrit dans le modèle de définition d'intron, par interaction du snRNP U2 avec le snRNP U1 lié au site 5' d'épissage en amont. Cependant, il a également été montré récemment que le complexe de définition d'exon contient aussi le tri-snRNP U4/U5-U6 et qu'il est capable de se convertir directement en complexe B (pour revue : Will and Lührmann, 2011). Les mécanismes permettant cette conversion restent aujourd'hui encore très mal connus, mais ils pourraient impliquer l'action d'élements de séquences *cis* régulatrices et des facteurs régulateurs de l'épissage, notamment la région riche en pyrimidines et la protéine PTB (Ram and Ast, 2007; Sharma *et al.*, 2008). Les réactions catalytiques se font par la suite par des interactions établies à travers l'intron, selon le modèle de définition de l'intron.

b. Le splicéosome mineur

L'assemblage du splicéosome mineur au niveau du pré-ARNm est très similaire à celui du splicéosome majeur. Néanmoins, des différences essentielles subsistent, notamment au niveau de la composition du splicéosome (Figure 24). En effet, bien que le splicéosome mineur U12-dépendant contienne également cinq snRNPs, U11, U12, U4atac, U6atac et U5, seul le snRNP U5 est partagé par le splicéosome majeur, les autres étant équivalents mais spécifiques du splicéosome mineur (pour revues : Patel and Steitz, 2003b; Turunen *et al.*, 2013; Will and Lührmann, 2005). Il existe également des différences essentielles dans les premières étapes d'assemblage du splicéosome, au moment de la reconnaissance des signaux d'épissage plutôt que pendant les réactions catalytiques. Tout d'abord, le complexe E n'existe pas dans le splicéosome mineur. En effet, le dimère snRNP U11/U12 (analogue du dimère U1/U2) se fixe de manière coopérative sur le site 5' d'épissage et la boite de branchement du pré-ARNm, ce qui correspond à la formation du complexe A, puis le tri-snRNP U4atac/U6atac.U5 est recruté pour former le complexe B. Après activation à la suite de réarrangements qui déstabilisent puis libèrent les snRNPs U11 et U4atac, le complexe B* actif catalyse alors la première réaction de trans-estérification et génère le complexe C-like qui peut alors procéder à la deuxième réaction de trans-estérification, avant dissociation complète du splicéosome, selon la même séquence d'événements observée dans l'activation du splicéosome majeur (pour revues : Patel and Steitz, 2003b; Turunen *et al.*, 2013; Will and Lührmann, 2005).

<u>Figure 24</u> : Etapes d'assemblage du splicéosome mineur et comparaison avec l'assemblage du splicéosome majeur (adapté de Will and Lührmann, 2005).



5) <u>Les éléments cis-régulateurs de l'épissage</u>

La plupart des transcrits humains, au même titre que les transcrits des mammifères, contiennent un très large nombre de séquences ressemblant aux sites consensus d'épissage. Pourtant bien plus nombreux que les sites d'épissage physiologiques, en particulier dans les grands introns, ces pseudosites d'épissage ne sont jamais utilisés dans des conditions normales (Krawczak *et al.*, 1992; Senapathy *et al.*, 1990; Sun and Chasin, 2000). En effet, ces pseudosites, aussi appelés sites cryptiques, sont efficacement ignorés pendant le processus d'épissage. D'autres signaux

auxiliaires présents le long de la séquence du pré-ARNm seraient donc impliqués dans le choix des sites d'épissage et seraient capables de distinguer les sites d'épissage réels des pseudosites et *vice versa* pour garantir la fidélité du processus d'épissage (Sun and Chasin, 2000). Il s'agit des éléments cis régulateurs d'épissage (SREs, *splicing regulatory elements* ; pour revues : Cartegni *et al.*, 2002; Chasin, 2007; Wang and Burge, 2008). Contrairement aux signaux consensus d'épissage (sites 5' et 3' d'épissage, le point de branchement et la région riche en pyrimidine) qui sont relativement bien définis, les éléments de régulation d'épissage restent très peu caractérisés. Ils sont généralement décrits comme de courtes séquences de 6 à 8 nucléotides constituant des sites de liaison pour des facteurs trans qui modifient l'activité du spliceosome en activant ou inhibant l'utilisation des sites d'épissage adjacents. Ils forment ainsi un code extrêmement complexe encore très partiellement caractérisé.

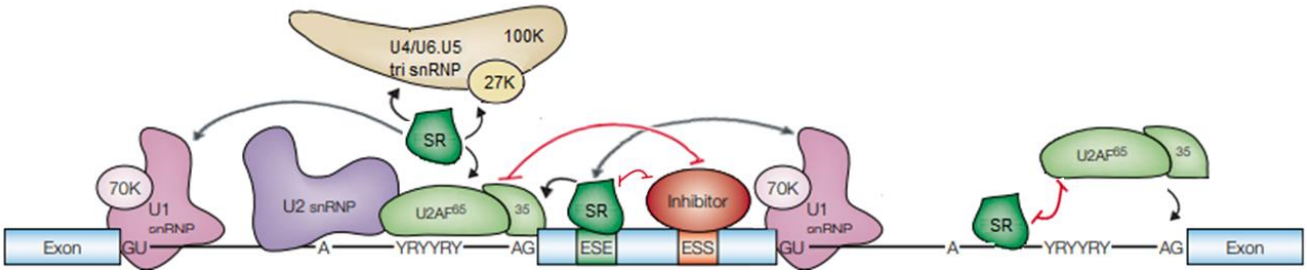
Les éléments de régulation de l'épissage agissent généralement à proximité, au niveau des sites d'épissage adjacents, localisés à 200-300 nucléotides. On parle alors d'éléments de régulation proximaux d'épissage. Mais les SREs peuvent moduler l'épissage et ce quelle que soit leur position dans le pré-ARNm par rapport à un exon donné et surtout leur distance par rapport à un site d'épissage donné. En effet, il a également été montré, plus récemment, que les éléments régulateurs pouvaient moduler l'utilisation de sites d'épissage localisés à une distance de plus d'un kilobase. Ces éléments de régulation d'épissage dits distaux pourraient avoir la même importance que des éléments situés à proximité d'un site d'épissage donné (pour revue : Fu and Ares, 2014). Les SRE peuvent également être classés selon leur localisation au niveau de la séquence du pré-ARNm ainsi que leur rôle activateur ou inhibiteur d'épissage par rapport aux sites d'épissage adjacents (Figure 21 ; pour revues : Cartegni *et al.*, 2002; Chasin, 2007; Wang and Burge, 2008). En effet, ces éléments peuvent être introniques (ISR, *intronic splicing regulatory elements*) ou exoniques (ESR, *exonic splicing regulatory elements*), et sont appelés *enhancers* ou *silencers* selon qu'ils stimulent ou inhibent l'inclusion de l'exon en reconnaissant des protéiques activatrices ou inhibitrices de l'épissage, respectivement. Plus particulièrement, les éléments activateurs localisés au niveau des exons (ESE, *exonic splicing enhancers*) ou des introns (ISE, *intronic splicing enhancers*) correspondent généralement à des séquences riches en purine reconnues par des protéines activatrices qui favorisent l'inclusion de l'exon. A l'inverse, les éléments inhibiteurs exoniques (ESS, *exonic splicing silencers*) ou introniques (ISS, *intronic splicing silencers*) sont reconnus par

des protéines inhibitrices favorisant l'exclusion de l'exon (pour revues : Cartegni *et al.*, 2002; Chasin, 2007; Wang and Burge, 2008).

a. Les éléments activateurs d'épissage et les protéines SR

Parmi les éléments activateurs de l'épissage, les éléments exoniques ou ESE, sont contenus dans la plupart des exons, et particulièrement au sein des exons alternatifs, où ils sont regroupés en grande densité autour des sites d'épissage (Figure 25 ; pour revues : Cartegni *et al.*, 2003; Fairbrother *et al.*, 2002; Jensen *et al.*, 2009; Wang and Burge, 2008). En effet, contrairement aux exons constitutifs, les exons alternatifs sont généralement plus courts et contiennent des sites d'épissage plus faibles, dont la reconnaissance par la machinerie d'épissage est favorisée par le recrutement des protéines SR aux niveaux des ESE. Les protéines SR constituent ainsi d'importants régulateurs de l'épissage altératifs (pour revue : Jeong, 2017). Ces éléments sont généralement reconnus par des membres de la famille des protéines riches en sérine/arginine (SR), notamment, par l'intermédiaire de leur(s) motif(s) de reconnaissance d'ARN à l'extrémité N-terminale (RRM, *RNA recognition motif*), qui assure la spécificité de liaison au substrat (Figures 25 et 26 ; Bourgeois *et al.*, 2004; Howard and Sanford, 2015; pour revue : Graveley, 2000). Les protéines peuvent ensuite interagir avec d'autres protéines SR, d'autres facteurs d'épissage ou des composants du splicéosome, par l'intermédiaire de leur domaine RS (riches en R/S, arginine et sérine) situé au niveau de l'extrémité C-terminale (pour revues : Graveley, 2000; Shepard and Hertel, 2009). Le domaine RS participe également à la régulation de l'activité et de la localisation des protéines SR via la phosphorylation intensive des résidus sérines par les SRPKs (*SR protein kinases* ; pour revues : Graveley, 2000; Shepard and Hertel, 2009).

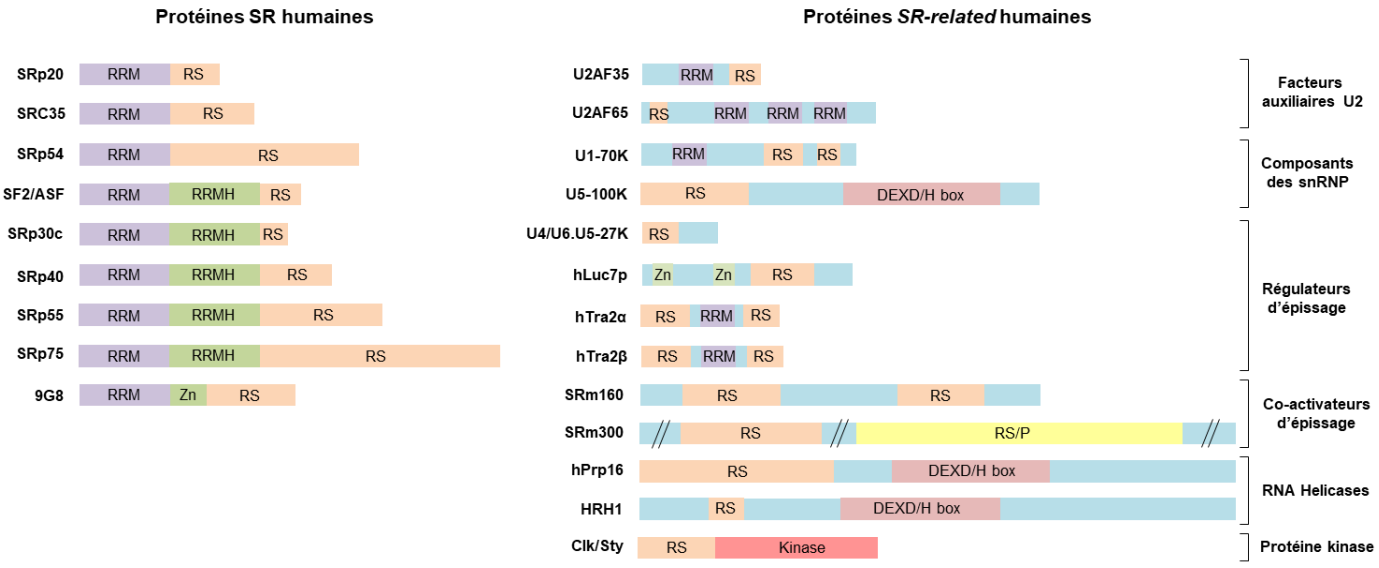
Figure 25 : Fonctions des protéines SR dans l'épissage de l'ARN messager (adapté de Cartegni *et al.*, 2002 et Shepard and Hertel, 2009).



Les protéines de la famille SR, caractérisées par la présence du domaine RS, regroupent une vingtaine de protéines, dont les protéines SFRS (*splicing factor, arginine/serine-rich*) notamment SF2/ASF, SC35, SRp20, SRp75, SRp40, SRp55, 9G8, SRp46, SRp30c, SRp38, SRp54, SRp35 (SFRS1 à SFRS12, selon la nomenclature actuelle de Manley and Krainer, 2010) et d'autres protéines additionnelles SR-*like*, SR-*related* ou SRrps (*SR-related proteins* ; Figure 26 ; pour revues : Graveley, 2000; Shepard and Hertel, 2009). Les protéines SR, exprimées de manière ubiquitaire, sont localisées pour la plupart dans le noyau où elles sont concentrées en *speckles*, domaines subnucléaires qui agissent comme un site de stockage pour certains facteurs d'épissage. Elles vont ensuite être recrutées et transportées vers les sites actifs de transcription, l'épissage étant un mécanisme co-transcriptionnel, où elles vont faciliter l'assemblage du splicéosome, et plus particulièrement la formation et la stabilisation du complexe E (Figure 25 ; pour revue : Graveley, 2000). En effet, après reconnaissance de manière spécifique puis interaction avec les ESE situés à proximité des sites d'épissage, les protéines SR peuvent alors recruter simultanément les facteurs U2AF35 et U1-70K au niveau du site 3' d'épissage en amont et du site 5' en aval, respectivement. De la même façon, elles vont pouvoir rapprocher les sites 5' et 3' d'épissage afin de faciliter l'élimination de l'intron. De plus, alors que les protéines hnRNP (*heterogeneous nuclear ribonucleoprotein particle*) recrutées au niveau des ESS vont bloquer la sélection du site 3' d'épissage par U2AF, les protéines SR recrutées au niveau des ESE et vont antagoniser l'action de ces répresseurs de l'épissage, favorisant ainsi la sélection des sites d'épissage (pour revue : Long and Caceres, 2009). Il est à noter que l'activité des protéines SR est dépendante du contexte de la séquence du pré-ARNm (Fu and Ares, 2014). En effet, celles-ci peuvent agir soit comme des facteurs activateurs d'épissage par leur fixation sur des séquences exoniques (ESE), soit comme inhibiteurs en se liant à des éléments introniques (ISS) (Figure 25 ; Shen and Mattox, 2012).

Quant aux éléments introniques activateurs de l'épissage ou ISE, ces derniers sont moins bien caractérisés que les éléments exoniques activateurs d'épissage. Néanmoins, quelques éléments ISE ont été particulièrement bien décrits notamment le triplet GGG ou la répétition Gn ($n \geq 3$), souvent distribués en *clusters* au niveau intronique afin de favoriser la reconnaissance des sites 5' et 3' d'épissage adjacents (McCullough and Berget, 1997, 2000) ou la répétition CA pour laquelle a été rapportée un rôle activateur de l'épissage sur les exons en amont (Hui *et al.*, 2005; Hung *et al.*, 2008).

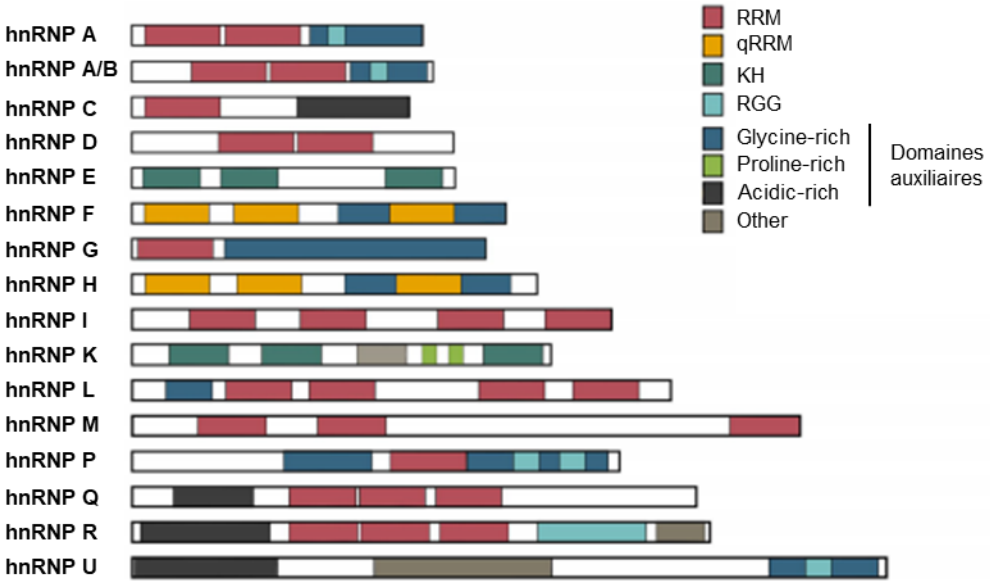
Figure 26 : Structure des protéines SR et SR-*related* humaines (adapté de Graveley *et al.*, 2000). RRM, *RNA recognition motif* ; RRMH, *RRM homology* ; Zn, *zinc knuckle* ; RS, *arginine/serine-rich domain* ; DEXD/H Box, *motif characteristic of RNA helicases*.



b. Les éléments inhibiteurs d'épissage et les facteurs hnRNP

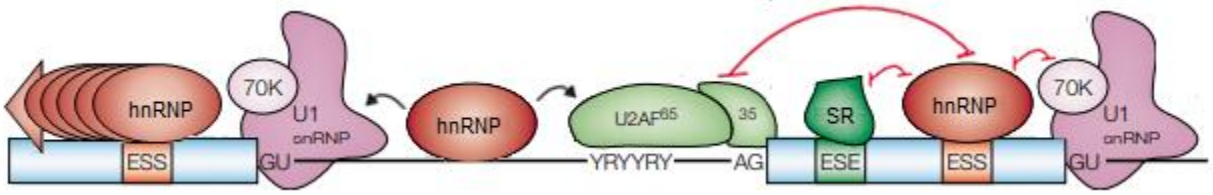
L'utilisation des sites 5' et 3' d'épissage peut être également influencée par des éléments inhibiteurs d'épissage, décrits comme des éléments particulièrement riches en UAG (Vorechovsky, 2010; Wang *et al.*, 2004). Ces éléments sont essentiellement reconnus par les protéines de la famille hnRNPs, par l'intermédiaire de leur(s) domaine(s) de fixation à l'ARN ou RBDs (*RNA-binding domains* ; Figure 27). Elles peuvent également contenir, pour certaines, des domaines d'inhibition d'épissage comme les motifs riches en glycine (pour revues : Pozzoli and Sironi, 2005; Wang and Burge, 2008). La famille des hnRNPs regroupent une vingtaine de membres nommés hnRNP A1 à U, exprimés de façon ubiquitaire, parmi lesquels les facteurs d'épissage hnRNP A1, hnRNP H et hnRNP I ou PTB (*polypyrimidine tract-binding protein*), dont l'activité est, comme les protéines SR, régulée par leur localisation intracellulaire et leur interaction avec d'autres protéines (Figure 27 ; pour revues : Geuens *et al.*, 2016; Jensen *et al.*, 2009; Martinez-Contreras *et al.*, 2007).

Figure 27 : Structure des protéines hnRNP (d'après Geuens *et al.*, 2016). RRM, *RNA recognition motif* ; qRRM, *quasi-RNA recognition motif* ; KH, *K-homology domain* ; RGG, R*NA-binding domain.*



Par leur fixation au niveau des sites ESS, les protéines hnRNP inhibent l'épissage en empêchant (i) la reconnaissance des sites d'épissage ainsi que le recrutement des facteurs d'épissage et (ii) en interférant avec les facteurs activateurs d'épissage (Figure 28 ; pour revues : Geuens *et al.*, 2016; Martinez-Contreras *et al.*, 2007). Par exemple, les protéines hnRNP A1/PTB et hnRNP L inhibent l'utilisation du site d'épissage en interférant avec la fixation des facteurs U2AF et snRNP U1 au niveau de la région riche en pyrimidine et du 5' d'épissage, respectivement, tandis que la protéine hnRNPA/B agit plutôt comme antagoniste des protéines SR (pour revue Fu and Ares, 2014). Bien souvent, la fixation de protéines hnRNPs favorisent le recrutement d'autres protéines hnRNPs grâce à des interactions protéines-protéines dites coopératives. L'ensemble des petits effets additifs exercés par les protéines hnRNPs vont créer une zone locale de répression qui va ensuite s'étendre petit à petit pour empêcher l'interaction des composants du splicéosome aux sites d'épissage et/ou la fixation des protéines SR sur les éléments activateurs (pour revue : Martinez-Contreras *et al.*, 2007). Cependant, comme les protéines SR, l'activité des protéines hnRNPs est dépendante du contexte génomique, parce que la liaison des facteurs hnRNPs à des ISS peut permettre, en fonction du contexte nucléotidique de leur fixation, à l'inverse une activation de l'épissage (pour revue : Fredericks *et al.*, 2015).

Figure 28 : Fonctions des protéines hnRNP dans l'épissage de l'ARN messager (adapté de Cartegni *et al.*, 2002 et Martinez-Contreras *et al.*, 2007).



6) Épissage alternatif de l'ARNm

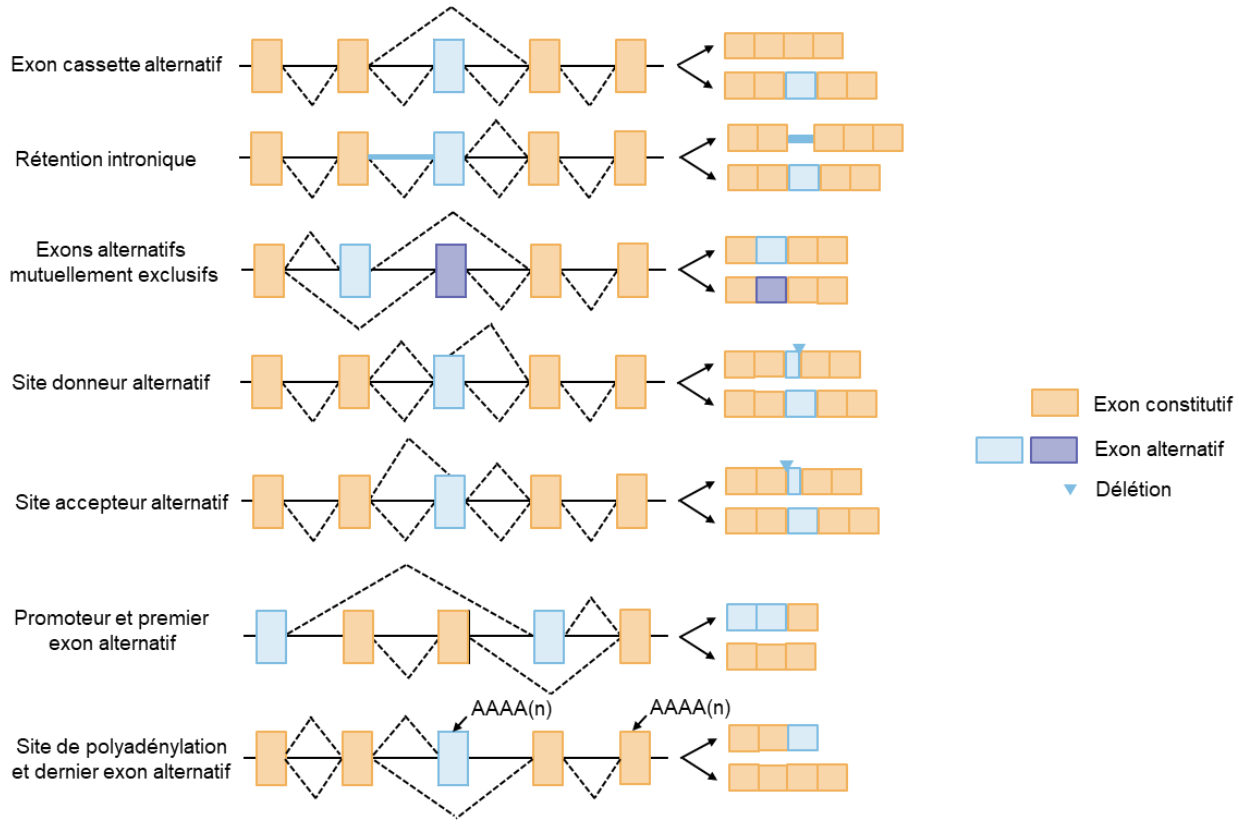
Proposé pour la première fois en 1978 par Walter Gilbert, le concept d'épissage alternatif est actuellement le mécanisme qui permet d'expliquer la contradiction entre le nombre de gènes codant pour des protéines (~25 000) et le nombre de protéines effectivement générées (>90 000) (Gilbert, 1978b; International Human Genome Sequencing Consortium, 2004), invalidant ainsi l'ancien dogme « un gène, une protéine, une fonction ». L'épissage alternatif est le mécanisme qui permet la production de différents transcrits ou isoformes à partir d'un seul et même gène, contribuant ainsi à la régulation qualitative de l'expression génique en fonction du tissu, du stade de développement ou en réponse à un signal donné (pour revue : Stamm *et al.*, 2005). En effet, si certains exons sont présents dans toutes les molécules d'ARNm, résultat d'un épissage constitutif, de nombreux exons sont alternativement épissés (pour revue : Wahl *et al.*, 2009). Des études transcriptomiques ont montré que plus de 95% des gènes humains sont soumis à un épissage alternatif (Pan *et al.*, 2008), en particulier dans le cerveau et les testicules, qui contiennent la plus grande proportion de transcrits alternatifs (Yeo *et al.*, 2004). Ces transcrits alternatifs produisant potentiellement des dizaines, voire des centaines de transcrits distincts, génèrent une grande diversité d'isoformes protéiques présentant des différences de stabilité, de localisation subcellulaire et de fonction (pour revue : Stamm *et al.*, 2005). Par exemple, un gène comme *DSCAM*, impliqué dans l'adhésion cellulaire, peut produire chez *Drosophila melanogaster*, dont le génome contient 15016 gènes, jusque 38016 transcrits différents (Schmucker *et al.*, 2000).

a. Les différents types d'épissage alternatif

Des analyses d'ESTs (*expressed sequence tag*) et de transcrits par l'utilisation de puces pangénomiques d'expression génique (*microarray*) ou par RNAseq ont rapporté l'existence de sept types d'événements d'épissage alternatif simple du pré-ARNm, non exclusifs et pouvant donner lieu, lorsque combinés, à des événements d'épissage dits complexes (pour revue : Blencowe, 2006). Il s'agit des évènements suivants : (i) le saut d'un exon cassette, événement d'épissage alternatif le plus simple, qui prédomine chez les mammifères, où un exon est entièrement exclu et absent dans le transcrit mature, (ii) l'utilisation d'un site 5' d'épissage alternatif, qui modifie la borne de l'exon à l'extrémité 3', conduisant à la délétion d'une partie de l'exon en 3' (ou alternativement à la rétention d'une partie de l'intron en aval), (iii) l'utilisation d'un site 3' d'épissage alternatif qui modifie la borne de l'exon à l'extrémité 5', conduisant à la délétion d'une partie de l'exon en 5' (ou alternativement à la rétention d'une partie de l'intron en aval), (iv) la rétention d'intron, événement d'épissage le plus rare chez les mammifères, consistant à l'inclusion d'un intron dans l'ARNm mature alors reconnu comme un exon, (v) des exons alternatifs mutuellement exclusifs lorsqu'un des deux exons est retenu dans l'ARNm mature et l'autre non, jamais les deux ensemble, (vi) un promoteur ou un premier exon alternatif et (vii) un site de polyadénylation ou dernier exon alternatif, qui ne correspondent pas à des événements d'épissage en soit puisque leur régulation dépend plutôt du choix des sites d'initiation et de terminaison de la transcription (Figure 29 ; pour revue : Blencowe, 2006).

Pour autant, l'ensemble de ces transcrits ne génèrent pas des protéines. En effet, il est estimé qu'environ un tiers des transcrits générés par épissage alternatif contiennent un PTC, cibles du NMD (*nonsense mediated mRNA decay* ; Chapitre IV ; Lewis *et al.*, 2003). Il s'agit d'un système de surveillance couplé à la traduction qui assure le contrôle qualité des ARNm et permet de réguler de manière quantitative l'expression des gènes en dégradant les transcrits porteurs d'un PTC. Si la plupart des transcrits porteurs d'un PTC représente un bruit de fond résultant d'erreur d'épissage aussi appelé épissage illégitime (*illegitimate splicing*), certains résultent d'un épissage alternatif régulé qui contribue à la régulation de l'expression des gènes selon le mécanisme RUST (*regulated unproductive splicing and translation* ; Chapitre V ; Losson and Lacroute, 1979; Peltz *et al.*, 1993; pour revue : Karousis *et al.*, 2016).

Figure 29 : Différents types d'épissage alternatifs (adapté de Blencowe, 2006).



b. Régulation de l'épissage alternatif

L'épissage alternatif s'explique par une modulation de la reconnaissance des sites d'épissage par le splicéosome, qui se traduit par l'utilisation préférentielle de sites 5' et 3' d'épissage au détriment d'autres sites d'épissage avec lesquels ils sont en compétition (Fu & Ares, 2014). Plusieurs facteurs intrinsèques, liés aux caractéristiques de l'exon, peuvent influencer l'utilisation des sites d'épissage, parmi lesquels la force des sites d'épissage, la taille de l'exon et des introns flanquants, la structure secondaire et le degré conservation (pour revue : Howard and Sanford, 2015). Il a d'ailleurs été rapporté que les exons alternatifs possèderaient des sites d'épissage plus faibles, et donc moins bien reconnus par la machinerie d'épissage, mais plus conservés que les exons constitutifs systématiquement retrouvés dans toutes les formes d'ARNm matures (pour revue : Fu and Ares, 2014). L'ensemble de ces caractéristiques permettent, au moins en partie, de distinguer les exons alternatifs des exons constitutifs (pour revue : Howard and Sanford, 2015).

De plus, le choix du site d'épissage est également contrôlé par les séquences *cis* régulatrices de l'épissage qui, en interagissant avec les protéines régulatrices, vont permettre d'influencer l'utilisation des sites d'épissage, en réprimant l'utilisation des sites d'épissage de référence ou en favorisant au contraire celle des sites d'épissage alternatifs en fonction qu'il s'agisse d'éléments inhibiteurs ou activateurs d'épissage, respectivement (pour revue : Lee and Rio, 2015). Bien que présents à la fois dans les exons constitutifs et alternatifs, les éléments *cis* régulateurs de type activateurs ont tendance à être déterminant dans l'épissage des exons constitutifs au sein desquels ils sont majoritairement concentrés, tandis que les éléments inhibiteurs sont relativement plus importants dans le contrôle de l'épissage alternatif et sont souvent retrouvés quant à eux dans les introns flanquants, et particulièrement autour des sites cryptiques (pour revues : Barash *et al.*, 2010; Wang and Burge, 2008). Par ailleurs, les éléments exoniques régulateurs d'épissage au même titre que les éléments introniques régulateurs d'épissage affichent un degré de conservation plus important dans les exons alternatifs (Goren *et al.*, 2006; pour revue : Jensen *et al.*, 2009), en particulier au niveau des régions qui entourent les sites 5' et 3' d'épissage ce qui souligne leur importance dans la définition des exons alternatifs (pour revue : Graveley, 2001).

Les séquences régulatrices sont reconnues par des facteurs de régulation d'épissage, dont l'expression est, comme l'épissage alternatif, un processus soumis à une régulation spatio-temporelle, qui peut varier en fonction des tissus et en fonction du stade de développement chez un même individu (Pan *et al.*, 2008; Merkin *et al.*, 2015; pour revue : Fredericks *et al.*, 2015). En effet, l'expression de nombreux facteurs de régulation d'épissage est tissu-spécifique, parmi lesquels NOVA, Fox, CELFs, MBNL, TIA et HuR (pour revue : Jensen *et al.*, 2009). Nova par exemple, spécifiquement exprimée dans le cerveau, permet de réguler un réseau de transcrits dont les fonctions interviennent dans le développement synaptique (pour revue : Jensen *et al.*, 2009). D'ailleurs, le fait que de nombreux facteurs d'épissage spécifiques des tissus agissent dans le cerveau est en accord avec l'observation selon laquelle le cerveau (avec les testicules) a la plus forte proportion de transcrits alternativement épissés (pour revue : Jensen *et al.*, 2009). Cette différence d'expression des facteurs de régulation d'épissage en fonction des tissus conditionne ainsi la différence d'expression des transcrits alternatifs (pour revue : Sveen *et al.*, 2016)

c. Epissage alternatif des gènes MMR et BRCA

Les travaux menés par Thompson *et al.* ont permis de référencer, à partir de la littérature et/ou des bases de données, l'ensemble des transcrits alternatifs identifiés dans les gènes MMR. En plus des transcrits pleine longueur, 30, 22, 4 et 9 transcrits alternatifs ont été identifiés pour les gènes *MLH1*, *MSH2*, *MSH6* et *PMS2*, respectivement. Certains de ces transcrits ont montré une différence d'expression en fonction des tissus (pour revue : Thompson *et al.*, 2015). Par exemple, quelques transcrits de *MLH1* ont été identifiés dans les lymphocytes mais pas dans le côlon (Genuardi *et al.*, 1998). Cependant, très peu d'études se sont focalisées sur la quantification des transcrits alternatifs des gènes MMR dans les tissus (Charbonnier *et al.*, 1995; Palmirotta *et al.*, 1998; Plaschke *et al.*, 1999; Takahashi and Nagai, 2009).

Au niveau fonctionnel, il est probable que la majorité des transcrits alternatifs des gènes MMR engendre des protéines non fonctionnelles à cause de la production de protéines tronquées, d'un effet dominant négatif de la protéine ou de la diminution de la quantité de transcrit pleine longueur (pour revue : Thompson *et al.*, 2015). En effet, 20% (n=13) des transcrits alternatifs identifiés sont prédits pour être entièrement non codants ou pour utiliser des sites d'initiation de la transcription ou de la traduction alternatifs et 57% (n=37) des transcrits potentiellement codants sont prédits comme étant cibles du NMD via l'introduction d'un codon stop prématuré (pour revue : Thompson *et al.*, 2015). Concernant les transcrits comportant des délétions internes en phase et codant potentiellement pour des isoformes protéiques fonctionnelles, les effets sur la fonction de très peu d'entre eux ont été appréhendés expérimentalement. En effet, des essais fonctionnels ont été réalisés *in vitro* pour seulement 3 isoformes protéiques de MLH1 : Glu578_Glu632del (Δ16), p.Glu633_Glu663del (Δ17) et p.Glu227_Ser295del (Δ9/10). Ces essais ont rapporté un défaut d'activité MMR pour l'ensemble des trois isoformes, par défaut d'interaction avec la protéine PMS2 pour Glu578_Glu632del (Δ16), et p.Glu633_Glu663del (Δ17), les exons 16 et 17 étant localisés au niveau des domaines d'interaction avec les protéines PMS2 et EXO1 ; et par défaut d'interaction avec Mutα pour p.Glu227_Ser295del (Δ9/10), les exons 9 et 10 étant localisés dans le domaine d'interaction avec l'hétérodimère Mutα (MSH2/MSH6) (Nyström-Lahti *et al.*, 1999; Raevaara *et al.*, 2005; Trojan *et al.*, 2002). A noter que cette dernière isoforme exerce même un effet dominant négatif dans les lignées cellulaires proficientes en activité MMR (Peasland *et al.*, 2010).

Concernant les gènes BRCA, les travaux menés par Colombo *et al.* et Fackenthal *et al.* ont permis d'identifier expérimentalement l'ensemble des transcrits alternatifs générés à partir des gènes *BRCA1* et *BRCA2*, respectivement. Ainsi, 63 et 24 transcrits alternatifs des gènes *BRCA1* et *BRCA2*, respectivement, ont été identifiés à ce jour dans des échantillons biologiques dérivés de sang d'individus contrôles (Colombo *et al.*, 2014; Fackenthal *et al.*, 2016). Ces échantillons biologiques correspondent à des lignées lymphoblastoïdes (LLBs), des leucocytes du sang périphérique stimulés en culture primaire (PBLs, *peripheral blood leucocytes*), des leucocytes du sang entier (LEUs, *whole blood leucocytes*) et/ou des cellules mononucléaires isolées (PBMCs, *percipheral blood mononuclear cells*). La majorité des transcrits, 62% (39 sur les 63 identifiés) pour *BRCA1* et 67% (16 sur les 24 identifiés) pour *BRCA2*, ont été détectés dans toutes les sources d'ARN et la plupart des transcrits 81% (51/63) et 79% (19/24) et ont été identifiés dans au moins 3 des 4 sources d'échantillons, sachant que la plupart des divergences peut être expliquée par la faible couverture de certains évènements d'épissage détectés. De manière intéressante, ces études ont également révélé que la très grande majorité des transcrits (*BRCA1* : 73%, 46/63 ; *BRCA2* : 48%, 10/21) détectés dans le sang ont également été détectés dans le tissu mammaire et aucun transcrit spécifique au tissu mammaire n'a été identifié. A noter que 100% (24/24) des transcrits alternatifs de *BRCA2* ont été identifiés dans au moins une des six lignées cellulaires dérivées du sein (MCF7, HCC1937, BT20, MCF10A, 184A1, 184B5), avec une fréquence variable. De plus, la quasi-totalité des transcrits alternatifs prédominants (BRCA1 : 8/10 et BRCA2 : 4/4) détectés à la fois dans les tissus sanguin et mammaire ont été quantifiés et montrent une expression similaire. Ces données suggèrent que les épissages alternatifs des gènes BRCA dans les tissus sanguin et mammaire sont similaires et supportent la relevance clinique des tests fonctionnels d'épissage *in vitro* à partir du sang des patients.

Comme pour les gènes MMR, la fonctionnalité des transcrits alternatifs BRCA et leur relevance biologique restent aujourd'hui très peu appréhendées. Si le caractère non fonctionnel de la plupart des transcrits BRCA (*BRCA1* : 36/63 et *BRCA2* : 19/24) parait évidente parce que ces transcrits sont non-codants suite à l'élimination du codon d'initiation de la traduction ou parce qu'ils induisent un décalage du cadre de lecture conduisant à l'introduction d'un PTC potentiellement cible du NMD, certains transcrits comportant des délétions internes en phase dans la région codante, des modifications des régions terminales non codantes (régions 5' et 3' UTR) ou des PTC présents au niveau du dernier et donc non ciblés par le NMD, pourraient générées des

isoformes fonctionnelles. A noter qu'un transcrit *BRCA1* particulier codant un peptide de 1399 acides aminés a été identifié. Ce transcrit nommé BRCA1-IRIS (IRIS, *in-frame reading of BRCA1 intron 11 splice variant*) contient l'ensemble des exons 1 à 11 ainsi qu'une petite partie de l'intron 11 (~102 nt), suite à la reconnaissance d'un signal de polyadénlyation localisé dans l'intron 11. (Colombo *et al.*, 2014; ElShamy and Livingston, 2004). Si certaines études ont démontré la perte de fonction des transcrits *BRCA1* Δ14-15 ; Δ17-19 ou *BRCA2* Δ3, d'autres ont rapporté que les protéines BRCA2 Δ12 ou BRCA1 Δ9-10 seraient fonctionnelles, au moins en partie (Biswas *et al.*, 2012; de la Hoya *et al.*, 2016a; Li *et al.*, 2009a; Sevcik *et al.*, 2012, 2013). Néanmoins, la fonctionnalité de la plupart des transcrits alternatifs potentiellement codants restent à analyser.

Les travaux récemment menés sur les gènes *MMR* et les gènes *BRCA* ont permis d'identifier et décrire la totalité des transcrits, constitutifs et alternatifs, générés physiologiquement à partir de ces gènes (Colombo *et al.*, 2014; Fackenthal *et al.*, 2016; Thompson *et al.*, 2015). Ces données constituent un outil précieux pour la conception et l'analyse (au niveau qualitatif et quantitatif, à la fois) des essais d'épissage *in vitro* effectués pour étudier la pathogénicité de variants identifiés dans le génome de patients. Par exemple, lors de l'analyse des profils d'épissage induits par une variation, les amorces pourront être ainsi positionnées de manière stratégique pour inclure ou exclure des événements d'épissage spécifiques en fonction de la position de la variation d'intérêt. De plus, ces travaux représentent un prérequis indispensable à la compréhension de la signification biologique des événements d'épissage alternatifs physiologiques identifiés dans ces gènes.

CHAPITRE IV : SURVEILLANCE ET CONTROLE QUALITE DES ARNm PAR LE NMD

Les ARNm sont souvent considérés comme des molécules intermédiaires entre l'ADN, support de l'information génétique et les protéines support de la fonction du gène. Il n'empêche qu'il est nécessaire d'assurer au niveau des ARNm le maintien de l'intégrité de l'information génétique. Afin de garantir la production d'ARNm matures et fonctionnels, les cellules ont développé divers mécanismes de contrôle de qualité des ARNm évitant ainsi la production de protéines aberrantes potentiellement délétères pour la cellule.

1) Mécanismes de surveillances des ARNm

Plusieurs mécanismes de surveillance des ARNm, actifs tout au long du *processing* des ARNm, de la synthèse à la traduction en passant par la maturation et l'export assurent une surveillance continue de l'intégrité des transcrits. Le premier mécanisme fait intervenir les exosomes nucléaires afin d'assurer une surveillance co-transcriptionnelle des transcrits. Ces exosomes sont impliqués dans la dégradation co-transcriptionelle des précurseurs des ARNm qui ne sont pas correctement maturés dans le noyau, en particulier les pré-ARNm non épissés (Bousquet-Antonelli *et al.*, 2000) ou avec une queue poly(A) aberrante (Hilleren *et al.*, 2001). Le second mécanisme garantit la surveillance des transcrits lors de l'export par le complexe du pore nucléaire. Seuls les ARNm correctement maturés et conformés seront exportés vers le cytoplasme (Dimaano and Ullman, 2004). Des mécanismes additionnels interviennent enfin de manière co-traductionnelle qui impliquent trois mécanismes de surveillance cytoplasmiques couplés à la traduction et dédiés à la surveillance des transcrits matures qui ont réussi à atteindre le cytoplasme (Shoemaker and Green, 2012). L'ensemble de ces mécanismes de surveillance utilisent des enzymes de la machinerie de dégradation physiologique des ARN normaux, mais utilisent des voies d'activation qui leur sont spécifiques afin d'assurer une dégradation très rapide des ARNm reconnus comme étant aberrants. Il s'agit du (i) *no-go decay*, garantissant la dégradation des transcrits comportant de fortes structures secondaires empêchant le déplacement du ribosome et bloquant la machinerie traductionnelle (Frischmeyer *et al.*, 2002; van Hoof *et al.*, 2002; Vasudevan

et al., 2002), (ii) du n*on-go decay*, assurant la dégradation des ARNm qui ne comportent pas de codon stop terminateur de la traduction (pour revues : Clement and Lykke-Andersen, 2006; Harigaya and Parker, 2010; Tollervey, 2006), et (iii) *du nonsense-mediated decay* (NMD) chargé de la dégradation des transcrits comportant un PTC (Losson and Lacroute, 1979; Peltz *et al.*, 1993; pour revue : Karousis *et al.*, 2016).

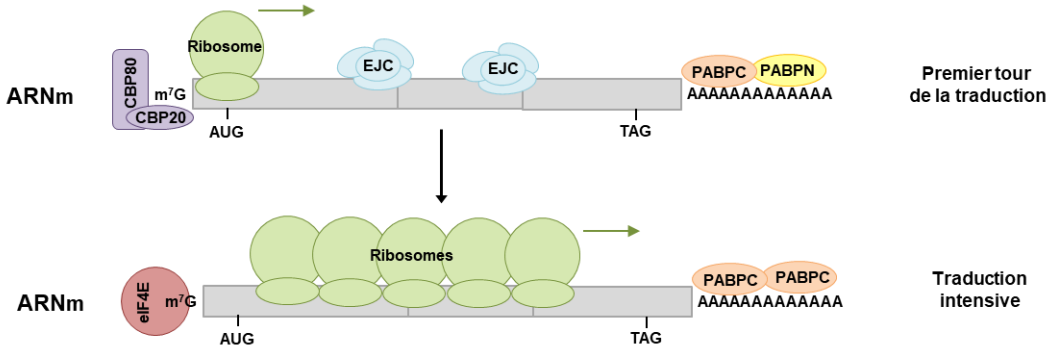
Il y a plus d'une quarantaine d'années, il a été découvert que des mutations terminant prématurément le cadre de lecture ouvert (ORF, *open reading frame*) réduisaient la demi-vie de ces ARNm (Losson and Lacroute, 1979). C'est en 1993 que Peltz et ses collaborateurs introduisent le terme de *nonsense mediated mRNA decay* pour décrire ce mécanisme complexe de surveillance dont le rôle est de coupler la reconnaissance d'un événement de terminaison prématurée de la traduction avec la dégradation spécifique et rapide de ces transcrits (Peltz *et al.*, 1993). Par conséquent, le NMD a été considéré comme un mécanisme de contrôle qualité des ARNm qui protège les cellules des effets délétères potentiels émanant de la génération de protéine tronquée dans leur extrémité C-terminale (pour revue : Karousis *et al.*, 2016).

2) Reconnaissance des codons stop prématurés (PTC)

Plusieurs événements peuvent induire l'introduction d'un PTC au niveau de la séquence d'un ARNm, soit au niveau de l'ADN par la survenue de mutations ponctuelles de type non-sens et d'insertions/délétions hors phase et de mutations d'épissage qui provoquent le décalage du cadre de lecture, soit au niveau de l'ARN à la suite d'erreur de l'ARN polymérase II pendant la transcription ou lors de la maturation du pré-ARNm par épissage aberrant ou alternatif (pour revue : Lejeune and Maquat, 2005). Bien que les mécanismes de reconnaissance d'un PTC ne soient pas encore totalement élucidés, plusieurs modèles, basés sur l'étude de différents organismes ont été proposés. Alors que ces modèles diffèrent sur plusieurs aspects, tous conviennent que le déclenchement du NMD dépend de la traduction (pour revue : Karousis *et al.*, 2016). En effet, lorsque la traduction est inhibée chimiquement au moyen d'inhibiteurs chimiques de la traduction (emetine, pyromycine, cycloheximide, l'anysomicine) ou physiquement, avec l'introduction d'une structure tige-boucle dans la partie 5'UTR de l'ARNm, ces ARNm sont stabilisés (Belgrader *et al.*, 1993; Carter *et al.*, 1995). Il a été montré que la reconnaissance des ARNm porteurs de PTC et leur dégradation se déroule au cours du premier tour de la traduction ce qui limite la synthèse de

protéines tronquées parce que la composition des ARNm qui subissent le premier tour de la traduction diffère de celle des ARNm qui subissent les autres tours de la traduction (Figure 30 ; Ishigaki *et al.*, 2001). Ces ARNm présentent avant traduction (i) l'hétérodimère CBP80-CBP20 au niveau de la coiffe, (ii) les complexes de jonction des exons (EJC, *exon junction complexe*) en amont des jonctions exon-exon et (iii) les protéines PABP nucléaires et cytoplasmiques (PABPN et PABPC) fixées sur la queue poly(A). Or, dans un ARNm normal, dans lequel le codon de terminaison est localisé sur le dernier exon, tous les EJC sont d'abord éliminés par le passage des ribosomes lors du premier tour de la traduction, de manière à ce qu'il ne reste plus d'EJC sur l'ARNm. Le complexe CBP80-CBP20 est remplacé par le facteur eIF4E, initiateur de la traduction. Seule la protéine PABPC reste associée à la queue poly(A) (pour revues : Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016).

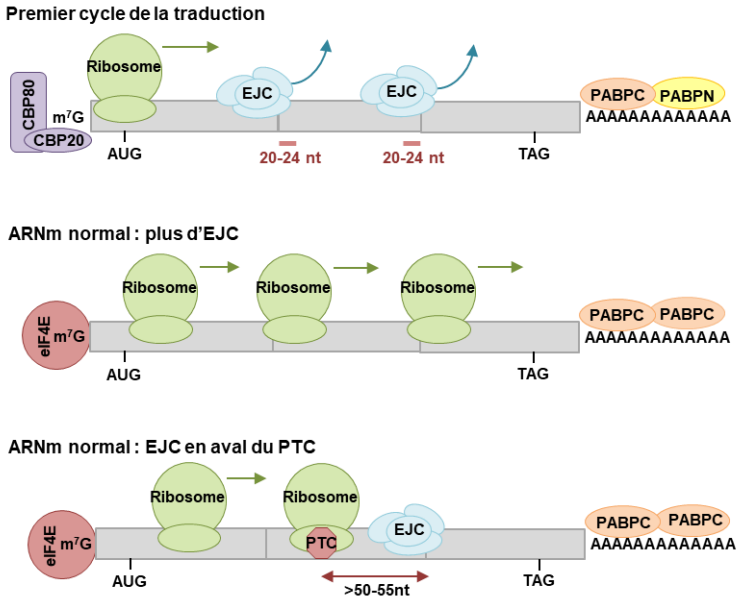
Figure 30 : Composition de l'ARNm avant et après le premier tour de traduction.



Le principal mécanisme de reconnaissance des PTC chez les mammifères fait intervenir l'EJC, un complexe protéique composé notamment des facteurs eIF4A3, MLN51, Y14 et Magoh formant le cœur de l'EJC, qui est déposé sur l'ARNm, 20-24 nucléotides en amont de chaque jonction exon-exon au cours de l'épissage (Figure 31 ; Le Hir *et al.*, 2000). Lorsque l'ARNm est porteur d'un PTC, le passage des ribosomes déplace les EJC qui se trouvent en amont du PTC. Cependant, lorsque les ribosomes rencontrent un PTC, ils s'arrêtent, et ce même si un ou plusieurs EJC sont encore présents en aval. C'est donc la présence de cet EJC, encore présent malgré le passage des ribosomes, qui indique la présence d'un PTC et déclenche le NMD qui aboutit à l'arrêt de la traduction et la dégradation des transcrits aberrants (pour revues : Brogna *et al.*, 2016; Chang

et al., 2007; Karousis *et al.*, 2016). Il est à noter que, selon la règle dite « des 50 à 55 nucléotides », seuls les codons stop localisés à plus de 50-55 nucléotides en amont d'une jonction exon-exon sont reconnus comme un codon stop prématuré (Nagy and Maquat, 1998). De plus, chez l'homme, seuls les ARNm porteurs d'un PTC nouvellement synthétisés sont soumis à la dégradation par le NMD, étant donné que seuls les ARNm porteurs de CBP80 et non eIF4E sont cibles du NMD *(Ishigaki et al.*, 2001). L'importance des EJCs pour le déclenchement du NMD chez les mammifères est soutenue par plusieurs observations : (i) la suppression des certaines protéines impliquées dans la formation des EJCs inhibe le NMD et inversement la fixation de ces protéines en aval d'un codon stop induit le NMD et (ii) des interactions directes entre des protéines de l'EJC et les facteurs du NMD ont été mises en évidence (Gehring *et al.*, 2003).

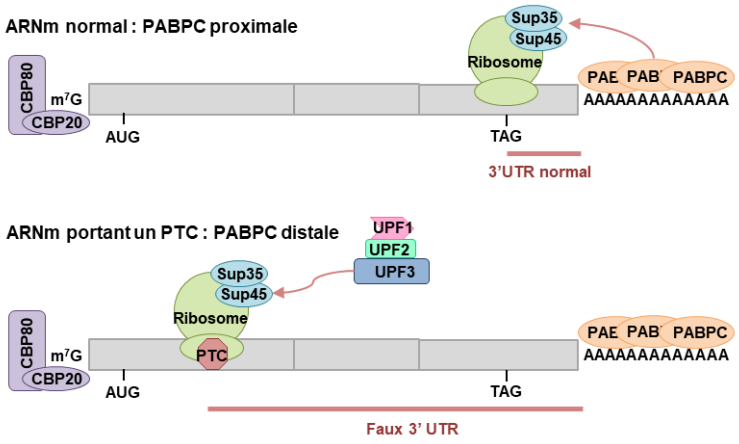
Figure 31 : Reconnaissance des PTC selon le modèle du complexe de jonction des exons (EJC) (adapté de Brogna *et al.*, 2016).



Chez d'autres espèces eucaryotes, en particulier D. *melanogaster*, C. *elegans* et S. *cerevisiae*, les EJCs ne sont pas indispensables à la reconnaissance d'un PTC (Gatfield *et al.*, 2003; Longman *et al.*, 2007). En effet, chez la levure moins de 5% des gènes possèdent un intron et les protéines de l'EJC sont absentes (Culbertson and Leeds, 2003). Ils existent donc d'autres mécanismes de

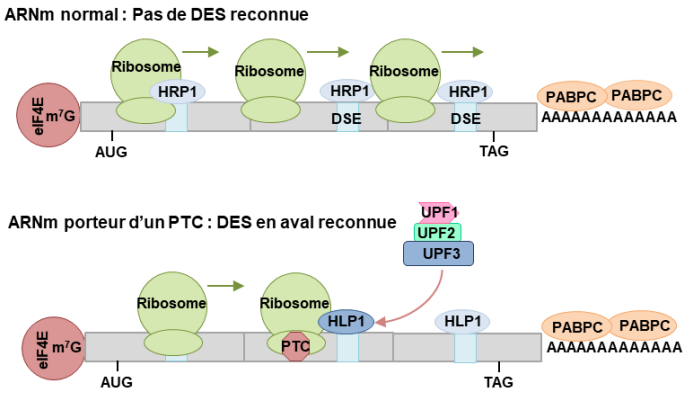
reconnaissance des PTC chez ces organismes, indépendants des EJCs, induits lorsqu'il n'y a pas d'EJC en aval du PTC (pour revue : Brogna *et al.*, 2016). Tandis qu'il est appelé modèle de « fausse 3'UTR » chez la levure, ce mécanisme est connu chez les mammifères sous le nom de NMD « *failsafe* », considéré comme un vestige du mécanisme de « fausse 3'UTR » (pour revue Brogna *et al.*, 2016). Il reste encore utilisé pour la dégradation sélective de certains transcrits dont les ARNm issus des gènes de *TPI* par exemple (Cheng *et al.*, 1994). Ce mécanisme fait intervenir la protéine PABPC (*cytoplasmic poly(A)-binding protein)*, présente au niveau de la queue poly(A) des ARNm (Figure 32). Il repose sur le fait que l'identification d'un codon stop comme étant un PTC serait dépendant de la longueur de la région 3'UTR, où un transcrit avec une distance anormalement longue entre le PTC et la queue poly(A) serait reconnu comme aberrant et dégradé. En effet, selon ce modèle, la terminaison normale de la traduction nécessite l'interaction entre le ribosome et la protéine PABPC. En revanche, lorsqu'un PTC est présent, celui-ci se retrouve entouré d'une « fausse 3'UTR » dans laquelle les ribosomes ne peuvent pas interagir avec PAPBC toujours présente au niveau de la réelle 3'UTR de l'ARNm. Les ribosomes restent ainsi bloqués sur le PTC et le complexe de surveillance NMD serait recruté afin de procéder à la dégradation de l'ARNm (Amrani *et al.*, 2004; pour revue : Brogna *et al.*, 2016) Ce modèle de « fausse 3'UTR » est compatible avec l'effet de polarité selon lequel le NMD est plus efficace lorsque le PTC est localisé le plus en 5' possible de la région codante, dans la première moitié (Losson and Lacroute, 1979).

Figure 32 : Reconnaissance des PTC selon le modèle de la fausse 3'UTR ou NMD « *failsafe* » (adapté de Brogna *et al.*, 2016).



Il a également été proposé un troisième modèle de reconnaissance des PTC, à partir d'études réalisées chez S. *cerevisiae* (pour revue : Brogna *et al.*, 2016). La reconnaissance des PTC serait liée selon ce modèle à la présence d'une séquence DES (*downstream sequence element*) en aval du codon stop, motif dégénéré présent en quelques copies le long de la séquence codante des ARNm qui serait d'autant plus probablement présente que le PTC serait proche de l'extrémité 5' du gène (Peltz *et al.*, 1993; pour revue : Brogna *et al.*, 2016). En effet, la présence de cette séquence dans les ARNm de la phosphoglycérate kinase (*PGK1*) induirait leur dégradation par le NMD, tandis que sa délétion permettrait de stabiliser les ARNm et d'abolir le NMD (Figure 33 ; Peltz *et al.*, 1993). Cette séquence permettrait le recrutement de facteurs stimulateurs du NMD, tel que la protéine Hrp1p (González *et al.*, 2000; Peltz *et al.*, 1993). Cette protéine peut interagir avec la protéine UPF1 du complexe de surveillance (UPF1-UPF2-UPF3) qui scanne la région 3'UTR en aval du codon stop à la rencontre d'un DES (González *et al.*, 2000; Peltz *et al.*, 1993). Si un DES est identifié, HRP1 interagit avec UPF1 pour identifier le NMD avant d'induire la dégradation de l'ARNm (González *et al.*, 2000; Peltz *et al.*, 1993). Cependant, il semblerait que cette voie d'activation du NMD serait très marginale. En effet, bien que les régions 3'UTR de la plupart des gènes de levure contiennent des séquences riches en A/U qui pourraient ressembler à des DES, des interactions DES-Hrp1p ont été identifiées uniquement pour l'ARNm *PGK1* (González *et al.*, 2000).

Figure 33 : Reconnaissance des PTC selon le modèle DSE (adapté de Brogna *et al.*, 2016).

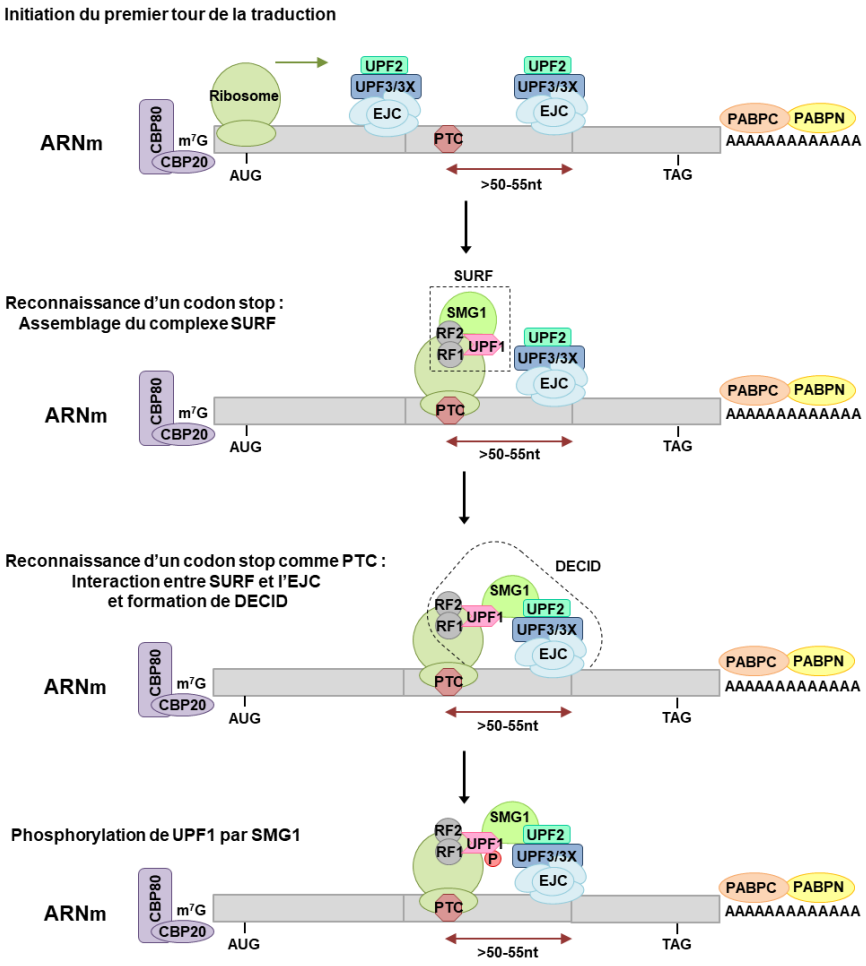


3) Assemblage du complexe de surveillance des PTC

Au cours du premier cycle de la traduction, les EJCs sont déplacés un à un par le ribosome jusqu'à ce qu'il rencontre un codon stop, reconnus par les facteurs de terminaison de la traduction eRF1 et eRF3 (*eukaryotic translation termination factor*) sur lequel ils sont recrutés, induisant ainsi la terminaison de la traduction (Figure 34 ; pour revues : Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016). Le complexe SURF (SMG1–Upf1–eRF1–eRF3 complex), comprenant les protéines SMG1, UPF1, eRF1 et eRF3, s'assemble alors au niveau de ce codon stop et ce indépendamment du complexe EJC (Kashima *et al.*, 2006). Lorsque ce codon stop est signalé comme étant un PTC par le complexe EJC présent en aval du PTC, le complexe SURF s'associe avec l'EJC par l'intermédiaire d'UPF3 lié à l'EJC et servant de pont pour UPF2 qui sert alors d'adaptateur entre l'interaction d'UPF3 avec UPF1 du complexe SURF. Ensemble, les complexes SURF et EJC forment le complexe DECID (*decay-inducing complex*) (pour revues : Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016). Ce complexe permet la phosphorylation de UPF1 par la protéine SMG1 (Kashima *et al.*, 2006).

Cet événement est crucial pour le devenir des ARNm puisqu'il déclenche les étapes subséquentes de dégradation (Kashima *et al.*, 2006). Pour autant, le mécanisme par lequel UPF1 phosphorylée induit les étapes subséquentes de dégradation reste encore mal connu. La phosphorylation d'UPF1 pourrait induire un remodelage du complexe de surveillance permettant le recrutement d'autres facteurs responsables de la déphosphorylation d'UPF1 et/ou de la dégradation de l'ARNm aberrant (Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016). En effet, UPF1 phosphorylée recrute le complexe SMG5-SMG7 interagissant avec la phosphatase PP2A, formant le complexe wee (WC), responsable de la déphosphorylation d'UPF1 *(Ohnishi et al.*, 2003). Cette étape est cruciale dans le mécanisme du NMD puisque la déphosphorylation d'UPF1 permet son recyclage. En effet, l'accumulation d'UPF1 phosphorylée par inhibition de l'interaction de hSMG5 avec hUPF1 via l'utilisation de mutants de SMG5 ou de la molécule inhibitrice NMDI1 entraine une inhibition du NMD (Durand *et al.*, 2007; Ohnishi *et al.*, 2003). De plus, la phosphorylation d'UPF1 induit un répression traductionnelle des ARNm porteurs d'un PTC de façon à ce que ces ARNm reconnus comme aberrants lors du premier cycle de la traduction ne peuvent subir d'autres cycles de traduction avant d'être dégradés (Isken *et al.*, 2008).

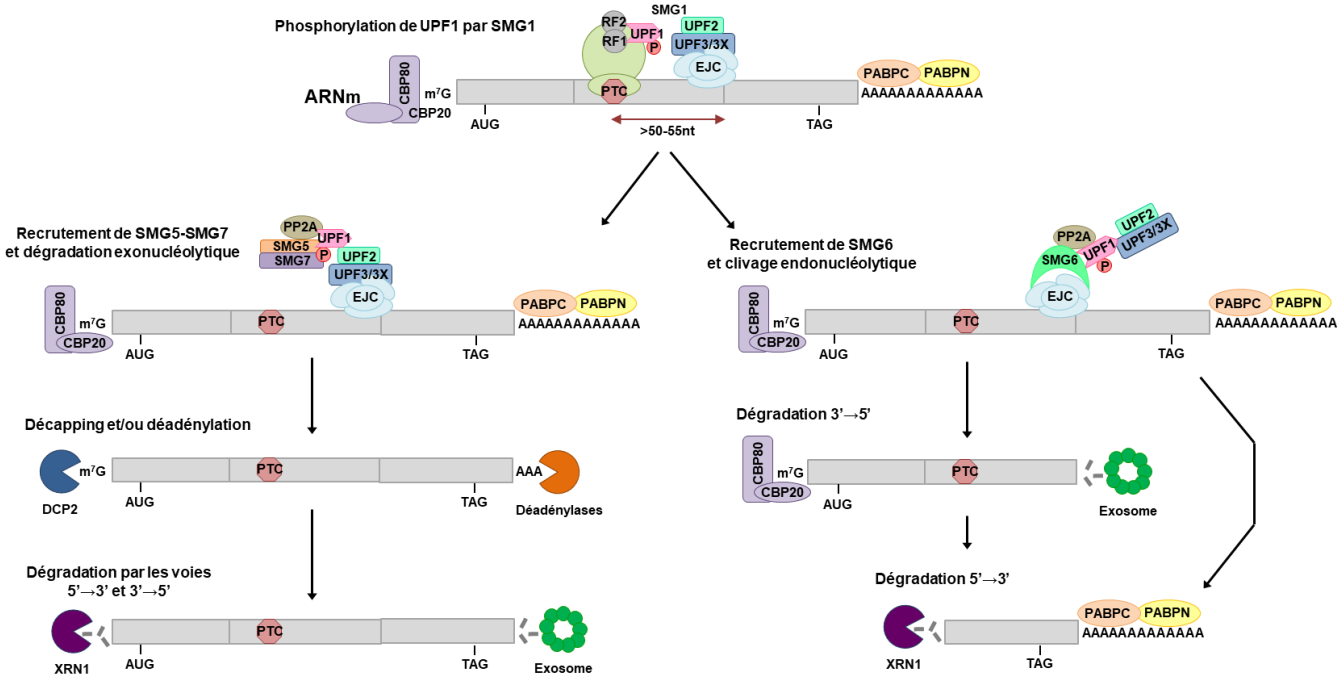
Figure 34 : Assemblage du complexe de surveillance (adapé de Kashima *et al.*, 2006).



4) Dégradation des ARNm aberrants porteurs d'un PTC

La dégradation des ARN soumis au NMD utilise les mêmes enzymes qui interviennent dans la dégradation physiologique des ARNm normaux (Figure 35 ; pour revues : Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016). Cependant, des voies d'activation spécifiques sont mises en jeu afin d'assurer une élimination rapide des transcrits aberrants. Il existerait deux voies possibles de dégradation des ARNm porteurs d'un PTC après la phosphorylation d'UPF1 (pour revues : Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016). La contribution individuelle de chacune de ces voies de dégradation reste inconnue ; il n'est pas exclu qu'elles interviennent de façon concomitante.

Figure 35 : Modèle des voies de dégradation des ARNm soumis au NMD.



Lorsqu'UPF1 est phosphorylée, celle-ci peut alors recruter l'hétérodimère SMG5-SMG7, qui peut alors recruter à son tour les enzymes de déadénylation, de decapping et enfin les exonucléases qui procèdent à la dégradation des ARNm porteurs de PTC (pour revues : Brogna *et al.*, 2016; Chang *et al.*, 2007; Karousis *et al.*, 2016). En effet, lorsque les ARNm doivent être dégradés, la configuration circulaire émanant de l'interaction, par l'intermédiaire du facteur EIF4G, entre l'extrémité 5' de l'ARNm (la coiffe) *via* le facteur eIF4E et l'extrémité 3' de l'ARNm (la queue poly(A)) *via* la protéine PABP, qui empêchait l'accès des facteurs de dégradation à l'ARNm, est déstabilisée (Gallie, 1998). La queue poly(A) devient accessible aux désanylases (complexes PAN2-PAN3 et CCR4-CAF1), qui vont procéder à son élimination ou déadénylation (Mitchell and Tollervey, 2000; Parker and Song, 2004). Deux voies de dégradation sont alors possibles : dans la première, un complexe de *decapping* (DCP1-DCP2) élimine la coiffe en 5' et expose le transcrit à la dégradation par les exoribonucléases 5'→3' ; dans la deuxième, le transcrit déadénylé est dégradé par l'extrémité 3' par les exoribonucléases 3'→5' de l'exosome (Mitchell and Tollervey, 2000; Parker and Song, 2004). UPF1 phosphorylée peut également recruter SMG6, une endonucléase capable de cliver l'ARNm à proximité du PTC. Ce clivage endonucléolytique génère deux fragments sont les extrémités non protégées sont accessibles aux exoribonucléases. Ils

peuvent ainsi être dégradés de manière subséquente par l'une des deux voies exonucléolytiques (Mitchell and Tollervey, 2000; Parker and Song, 2004).

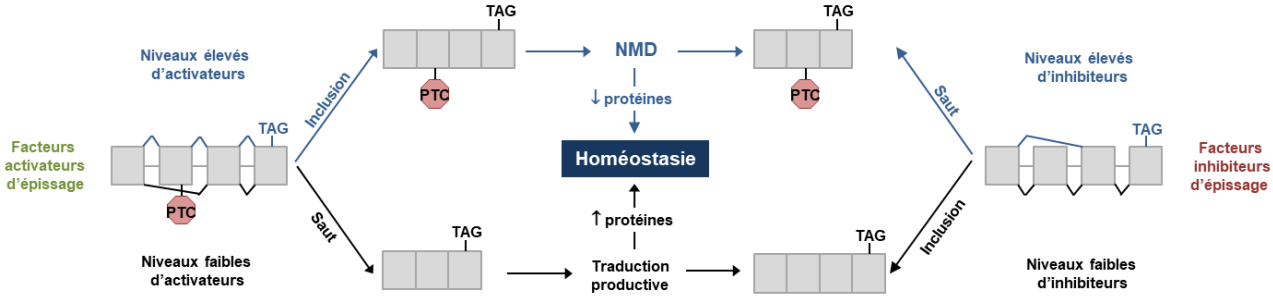
5) Importances physiologiques du NMD

L'implication du NMD a été démontrée dans de nombreux processus cellulaires centraux indispensables à l'activité de la cellule, en particulier le développement embryonnaire (pour revue : Karousis *et al.*, 2016). En effet, les facteurs essentiels à l'activité du NMD sont requis pour la viabilité de nombreux organismes (Kerényi *et al.*, 2008). Chez la souris, par exemple, l'inhibition de l'expression des facteurs UPF1, UPF2 ou SMG1 par knock-out induit la mort très tôt pendant le développement embryonnaire *(McIlwain et al., 2010; Medghalchi et al., 2001; Weischenfeldt et al.*, 2008). Cependant, ces effets pourraient être également dus à (i) la dérégulation de gènes dont l'expression est régulée au moins en partie par le NMD et (ii) d'autres processus cellulaires d'importance vitale dans lesquels le NMD est directement impliqué (modulation de la réponse au stress, développement cérébral et lymphocytaire) ou indirectement, au travers des protéines du NDM impliquées dans d'autres processus (maintien des télomères, progression du cycle cellulaire, métabolisme du virus HIV-1 parmi tant d'autres) (pour revue : Karousis *et al.*, 2016).

L'une des sources les plus importantes qui génère des transcrits non-sens porteurs d'un PTC est l'épissage alternatif de l'ARN pré-messager (pour revue : Lejeune and Maquat, 2005). En effet, la plupart des gènes chez l'homme, plus de 95% des gènes, subissent un épissage alternatif (Pan *et al.*, 2008). Cependant, tous les transcrits alternatifs se sont pas pour autant traduits en protéines. En effet, des analyses bioinformatiques menées sur des banques EST (*expressed sequence tag*) ont permis d'estimer qu'environ un tiers des événements d'épissage alternatif génère des transcrits porteurs d'un PTC, cibles du NMD (Lewis *et al.*, 2003). Ce mécanisme appelé RUST et correspondant à un couplage entre l'épissage et le NMD constitue un mécanisme de régulation post-transcriptionnelle de l'expression des gènes (Lewis *et al.*, 2003). Il est également retrouvé sous le terme de AS-NMD (*alternative splicing coupled to nonsense-mediated mRNA decay*) (pour revues : da Costa *et al.*, 2017; Karousis *et al.*, 2016). En effet, il a été montré que le NMD contrôle les niveaux d'expression physiologiques de 10 à 20% de l'ensemble des transcrits (Mendell *et al.*, 2004). L'existence d'un tel mécanisme de régulation a été mise en évidence pour la régulation de l'expression de facteurs d'épissage, en particulier des protéines SR tel que SC35 (Figure 36).

L'expression de SC35 est régulée par une boucle de régulation qui implique un épissage alternatif couplé à la dégradation par le NMD. En effet, lorsque la protéine SC35 est très abondante dans les cellules, elle inhibe sa propre expression en stimulant l'inclusion d'exons contenant un codon stop, conduisant ainsi à la production de variants d'épissage qui vont être cibles de la dégradation par le NMD. Des niveaux de protéine SC35 faibles font favoriser à l'inverse l'exclusion de ces mêmes exons et conduire à la production d'un ARNm traduit en une protéine fonctionnelle (Lareau *et al.*, 2007; Sureau *et al.*, 2001; pour revue : Karousis *et al.*, 2016).

<u>Figure 36 :</u> Régulation homéostatique de l'expression des facteurs régulateurs de l'épissage par le mécanisme d'épissage alternatif couplé au NMD (d'après Karousis *et al.*, 2016).



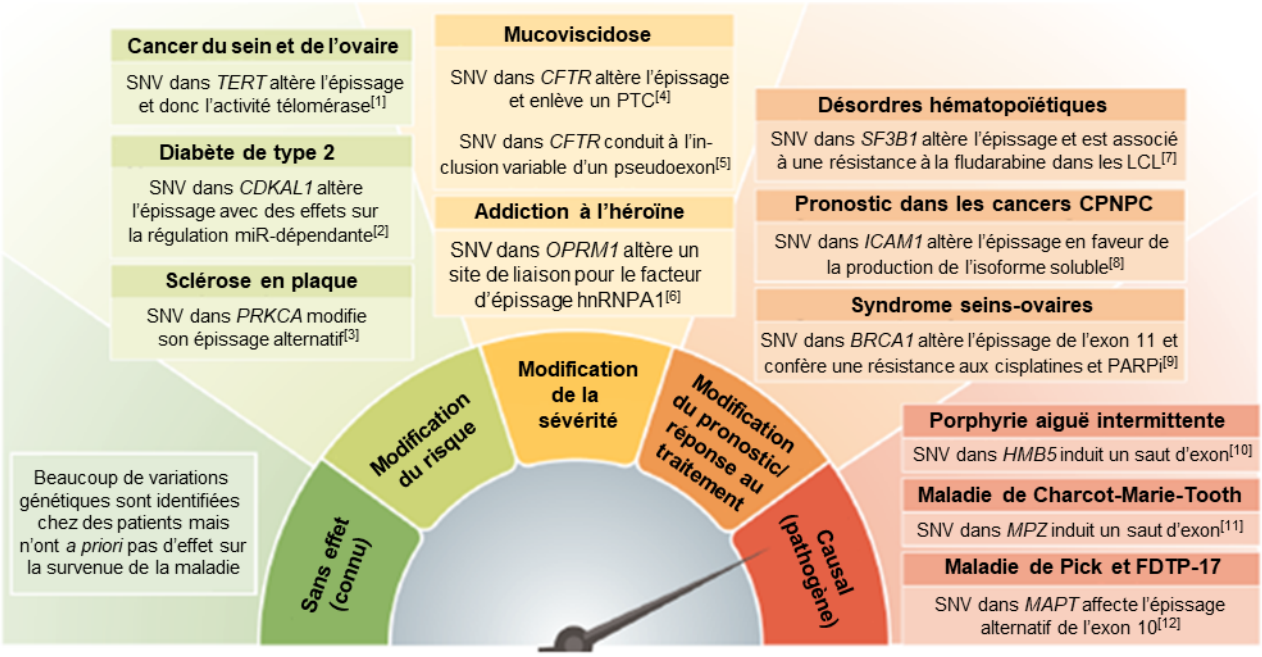
Il a également été montré que le NMD intervenait plus spécifiquement dans la régulation de l'expression des gènes essentiels au développement lymphocytaire, au niveau du processus de maturation des lymphocytes B et T. En effet, les gènes codant pour les TCR et les immunoglobulines, formés de nombreux segments V (variable), J (jonction) et D (diversité) subissent des réarrangements programmés impliquant des enzymes de recombinaison spécifiques, dont le rôle est d'assembler ces segments par un processus de recombinaison aléatoire, appelé recombinaison V(D)J. Puisqu'il existe de multiples segments, ce processus génère une très grande diversité du répertoire de lymphocytes T et B. A cela s'ajoute l'implication de l'enzyme transférase terminale, qui par l'ajout de nucléotides aux jonctions entre les segments, accroit encore cette diversité. Deux fois sur trois, ces réarrangements sont non fonctionnels par modification du cadre de lecture et apparition d'un PTC. Le NMD, qui va cibler ces réarrangements non fonctionnels porteurs de PTC, joue alors un rôle essentiel dans le maintien d'un répertoire de récepteurs B et T fonctionnels via la dégradation des transcrits aberrants (Li and Wilkinson, 1998).

CHAPITRE V : IMPLICATION DE L'EPISSAGE DANS DES MALADIES

L'épissage dépend d'un code complexe d'éléments *cis* présents le long de la séquence du pré-ARNm et d'un réseau important de facteurs *trans* qui interagissent de manière spécifique avec l'ARN mais également entre eux (pour revue : Cooper *et al.*, 2009). L'ensemble de ces éléments *cis* et de ces facteurs *trans* d'épissage peuvent être altérés par des mutations pouvant être à l'origine des maladies génétiques. En effet, ces mutations peuvent conduire à différents événements d'épissage alors considérés comme des anomalies de l'épissage : un saut d'exon, une augmentation de l'inclusion ou de l'exclusion d'un exon alternatif, une rétention d'intron ou une utilisation/destruction d'un site d'épissage cryptique ou alternatif (Figure 29 ; pour revues : Caminsky *et al.*, 2014; Ward and Cooper, 2010). Si la plupart de ces événements d'épissage conduisent à l'introduction d'un PTC via un décalage du cadre de lecture et à donc à la dégradation des transcrits, certains permettent la production d'une protéine tronquée, la traduction d'une région non codante ou l'altération de la composition de la protéine (pour revue : Lykke-Andersen and Jensen, 2015). A cela s'ajoute une altération du ratio des transcrits alternatifs qui peut survenir lorsque le gène est épissé de manière alternative (pour revue : Ward and Cooper, 2010).

De nombreuses maladies monogéniques et multifactorielles, allant des troubles neurologiques aux syndromes métaboliques, des maladies myogéniques aux maladies cardiovasculaires en passant par les cancers, sont associées à des anomalies de l'épissage (pour revue : Daguenet *et al.*, 2015). Ces dernières peuvent être directement impliquées dans l'étiologie de la maladie ou avoir des contributions plus subtiles dans la détermination de la susceptibilité, dans la progression ou dans la modulation de la sévérité de la maladie (Figure 37 ; pour revues : Cooper *et al.*, 2009; Ward and Cooper, 2010). De plus, les profils d'épissage associés à certaines maladies, en particulier les cancers, constituent de véritables signatures spécifiques de ces maladies, pouvant être utilisées comme marqueurs diagnostics et peuvent être indicatives d'une résistance thérapeutique. Par ailleurs, certaines de ces altérations ont émergé en tant que cibles thérapeutiques potentielles et permettent d'orienter la sélection de la stratégie thérapeutique (pour revue : Singh and Eyras, 2017).

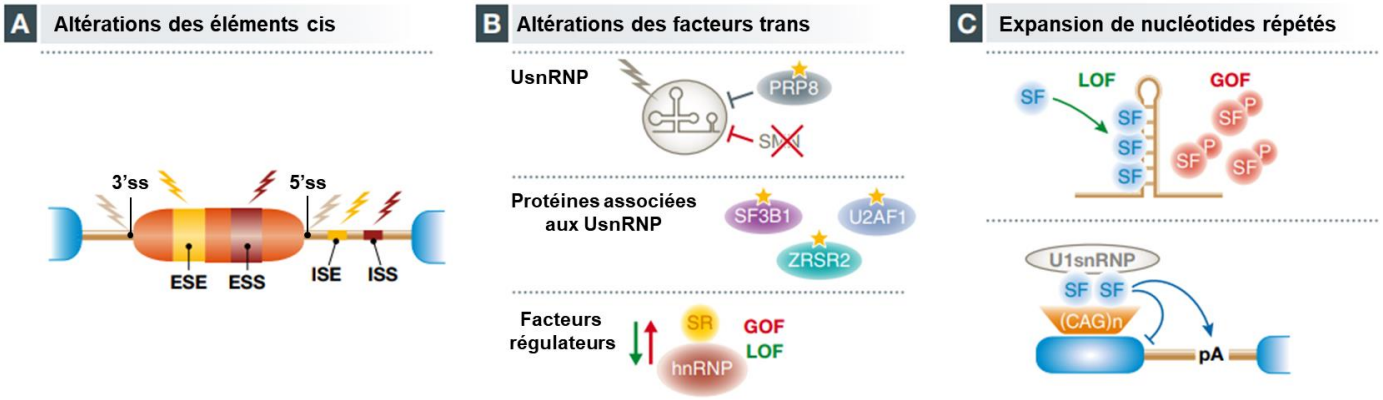
Figure 37 : Implication des variations génétiques qui affectent l'épissage des pré-ARNm dans la survenue des maladies (adapté de Manning and Cooper, 2017). [1] Paraboschi *et al.*, 2014, [2] Zhou *et al.*, 2014, [3] Bojesen *et al.*, 2013, [4] Chiba-Falek *et al.*, 1998 [5] Hinzpeter, *et al.*, 2010, [6] Xu *et al.*, 2014, [7] Papaemmanuil *et al.*, 2011, [8] Thanopoulou *et al.*, 2012, [9] Wang *et al.*, 2017 [10] Llewellyn *et al.*, 1996, [11] Corrado *et al.*, 2016, [12] Liu and Gong, 2008



1) Mécanismes d'altération d'épissage dans les maladies

La survenue d'une anomalie d'épissage associée à une maladie génétique peut être liée à trois mécanismes : (i) une mutation dans des éléments *cis* d'épissage, (ii) une mutation dans des facteurs *trans* d'épissage, ou (iii) des déséquilibres de la stœchiométrie des facteurs d'épissage suite à leur séquestration par des séquences répétées au niveau des ARN « toxiques » (Figure 38 ; pour revues : Daguenet *et al.*, 2015; Fredericks *et al.*, 2015; Ward and Cooper, 2010).

Figure 38 : Mécanismes d'altération de l'épissage dans les maladies génétiques (d'après Daguenet *et al.*, 2015). (A) Altérations des éléments *cis* d'épissage. (B) Altérations des facteurs *trans* d'épissage. (C) Mécanisme de l'ARN toxique. Par expansion de nucléotides répétés. SF, *splicing factor*; LOF, *loss of function* ; GOF, *gain of function*.



a. Altérations des éléments cis d'épissage (cis-acting mutation)

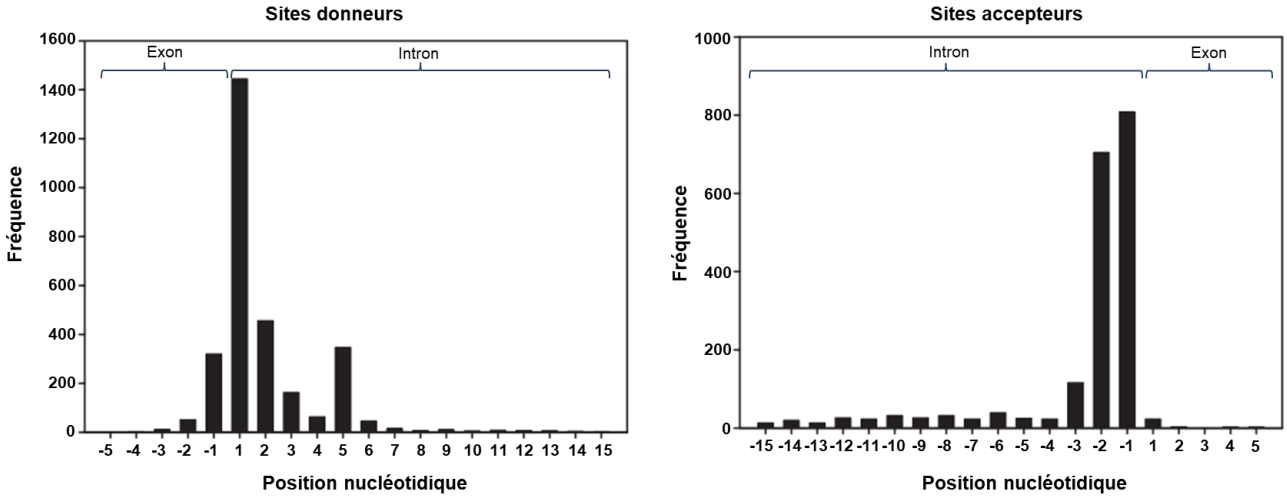
Il est actuellement estimé qu'environ 9,4% du nombre total des variations nucléotidiques répertoriées dans la base de données internationale HGMD (*human gene mutation database*) et 13,8% des substitutions ponctuelles sont des mutations d'épissage (Krawczak *et al.*, 2007; Stenson *et al.*, 2014; pour revue : Fredericks *et al.*, 2015). Les mutations dites d'épissage correspondent à des variations nucléotidiques qui altèrent l'épissage via la modification des éléments *cis*. Ces derniers regroupent les principaux signaux d'épissage (les sites 5' et 3' d'épissage, le point de branchement et la région riche en pyrimidine) ainsi que les éléments régulateurs de l'épissage. Ce type d'altération, fréquent dans les maladies héréditaires et le cancer, peut affecter l'épissage constitutif et alternatif (pour revue : Daguenet *et al.*, 2015).

- Principaux signaux d'épissage

De nombreuses mutations affectant les signaux principaux d'épissage et particulièrement les sites 5' et 3' d'épissage sont associées à des maladies génétiques (pour revue : Anna and Monika, 2018). La plupart des mutations qui affectent les sites 5' et 3' d'épissage touchent principalement les positions les plus conservées des sites d'épissage, à savoir les dinucléotides GU/AG au niveau intronique (pour revue : Fredericks *et al.*, 2015). Dans la base de données HGMD, 64% des mutations d'épissage affectant les sites donneurs d'épissage se situent au niveau du dinucléotide

GU et 77% de celles qui affectent les sites accepteurs d'épissage sont localisées au niveau du dinucléotide AG (Figure 39 ; Krawczak *et al.*, 2007). En plus de ces dinucléotides, d'autres positions au niveau des sites donneurs et accepteurs d'épissage sont également altérées par des mutations. Il s'agit des deux derniers nucléotides exoniques au niveau du site donneur (en position -2 et -1 par rapport au site donneur d'épissage) et des nucléotides situés en position +3 jusqu'à +6 de l'intron, et des nucléotides en positon -3 de l'intron au niveau site accepteur d'épissage (Figure 39 ; Krawczak *et al.*, 2007).

Figure 39 : Distribution des substitutions ponctuelles causales, répertoriées au niveau des sites donneurs et accepteurs d'épissage dans la base de données HGMD en 2006 (d'après Krawczak *et al.*, 2007).



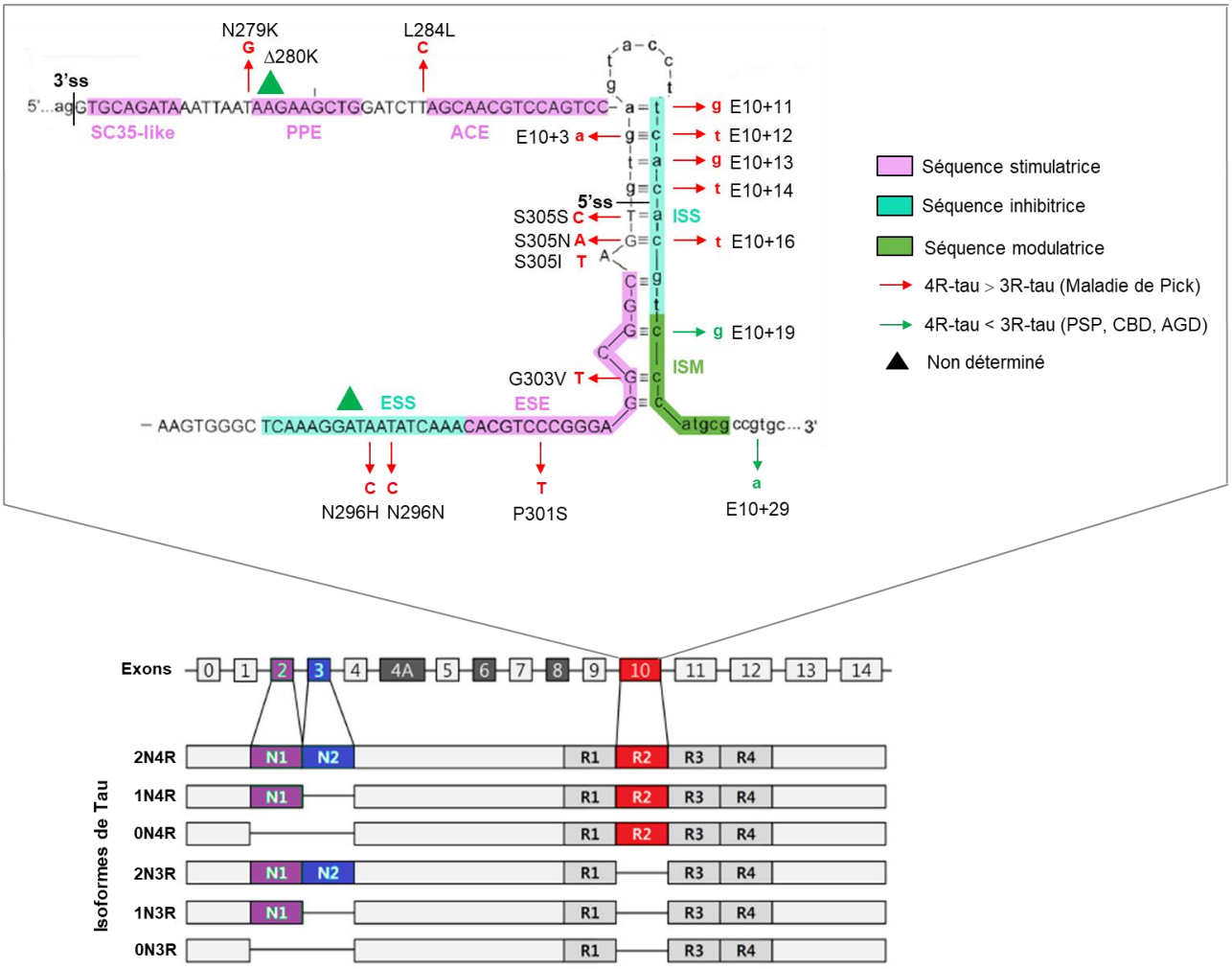
En général, lorsqu'un site d'épissage naturel est altéré ou détruit par une mutation, le splicéosome a tendance à utiliser le site d'épissage accessible le plus proche (Krawczak *et al.*, 2007). Dans la majorité des cas, où le site d'épissage utilisé est le site naturel de l'exon voisin, cela conduit à un saut d'exon. Par exemple, la variation *IKBKAP* c.2204+6T>C (IVS20+6T>C), retrouvée dans au moins 98% des allèles à l'origine de la dysautonomie familiale induit un saut de l'exon 20 via l'altération du site donneur d'épissage (pour revue : Rubin and Anderson, 2008). Cette anomalie d'épissage induit un décalage du cadre de lecture et conduit à l'introduction d'un PTC. Le transcrit résultant du saut de l'exon 20 est alors dégradé par le système NMD, diminuant ainsi le taux de protéine IKAP fonctionnelle (Slaugenhaupt *et al.*, 2001; pour revue : Ward and

Cooper, 2010). De même, des anomalies touchant des sites d'épissage U12 ont été également décrites dans les gènes *LKB1* et *SEDL* impliqués, respectivement, dans le syndrome de Peutz-Jeghers et dans la dysplasie spondyloépiphysaire tardive, comme à l'origine d'un saut d'exon (pour revue : Turunen *et al.*, 2013). Cependant, lorsque le splicéosome reconnait des sites d'épissage cryptiques ou alternatifs plus proches par rapport au site d'épissage altéré, cela peut générer d'autres types d'événements tels que la rétention d'une partie de l'intron ou la délétion d'une partie de l'exon. De même, certaines variations peuvent être responsables non pas de la destruction du site d'épissage naturel mais de la création d'un site d'épissage *de novo* ou de l'activation d'un site d'épissage cryptique qui vont entrer en compétition avec le site d'épissage naturel. Ces variations vont être à l'origine de la délétion d'une partie de l'exon ou la rétention d'une partie de l'intron contigu ou pas à un exon (création d'un pseudoexon) (pour revues : Dhir and Buratti, 2010; Vaz-Drago *et al.*, 2017). En effet, la variation *CFTR* c.1680-886A>G, identifiée chez des patients atteints de mucoviscidose, est responsable de la création d'un site donneur d'épissage dans l'intron 11 du gène *CFTR*, utilisé en combinaison avec un site accepteur présent 49 nucléotides en amont. Cette variation entraîne ainsi à la rétention d'un pseudoexon de 49 nucléotides entre les exons 11 et 12 du gène *CFTR*, aboutissant à un décalage du cadre de lecture et l'introduction d'un PTC dans les transcrits alors dégradés par le NMD (Chillón *et al.*, 1995).

En altérant les sites naturels d'épissage, d'autres variations, parmi lesquelles les variations localisées à proximité des sites d'épissage de l'exon 10 du gène *MAPT*, peuvent altérer le ratio des transcrits alternatifs et la stœchiométrie des isoformes protéiques résultantes. Le gène *MAPT*, codant pour la protéine tau impliquée dans certain nombre de maladies neurodégénératives appelées taupathies, est épissé de manière alternative (pour revue : Liu and Gong, 2008). En effet, l'épissage alternatif de l'exon 10 de *MAPT* génère deux isoformes contenant 3 ou 4 motifs de liaison aux microtubules (*microtubule-binding repeats*), appelées respectivement forme 3R-tau et 4R-tau, selon que l'exon 10 soit exclus ou inclus par la machinerie d'épissage (Figure 40). A l'état normal, le niveau d'expression des deux isoformes est similaire. Cependant, dans les taupathies dans lesquelles sont impliquées des variations situées à proximité des sites d'épissage de l'exon 10 de *MAPT* (i.e. IVS10+13A>G), le ratio 3R-tau/4R-tau dans le cerveau est déséquilibré *(Hutton et al.*, 1998). En effet, la variation IVS10+13A>G, située en dehors du site donneur d'épissage, permet l'ouverture d'une structure en tige-boucle de l'ARN qui, à l'état normal, séquestre partiellement le site 5' d'épissage et empêche l'inclusion totale de l'exon (Figure 40 ; Grover *et*

al., 1999; Hutton *et al.*, 1998). L'augmentation du taux de 4R-tau, via une augmentation de l'inclusion de l'exon, est à l'origine de la maladie neurodégénérative FDTP-17 (*frontotemporal dementia and Parkinsonism linked to chromosome 17*). A noter qu'une augmentation du saut de l'exon 10 de *MAPT* causée par des mutations dans cet exon est à l'origine de la maladie neurodégénérative de Pick (pour revue : Liu and Gong, 2008).

Figure 40 : Régulation de l'épissage alternatif de l'exon 10 de la protéine Tau, impliqué dans des nombreuses maladies neurodégénérative. Six isoformes de Tau sont exprimées dans le cerveau humain grâce à différentes combinaisons des exons 2, 3 et/ou 10 par épissage alternatif. L'exon 10 code pour le deuxième domaine de liaison aux microtubules (R2). Selon la présence de ce domaine R2, la protéine devient la protéine 3R-tau (exclusion de l'exon) ou 4R-tau (inclusion de l'exon). De nombreuses mutations, via une altération des sites d'épissage, de la structure secondaire ou des séquences régulatrice de l'épissage affectent le ratio 3R/4R conduisant alors à des pathologies neurodégénératives (d'après Liu *et al.*, 2010 ; Park *et al.*, 2016).



En plus des sites d'épissage, les points de branchement et la région riche en pyrimidine peuvent également être affectés par des mutations (pour revue : Anna and Monika, 2018). L'une des premières altérations d'un point de branchement a été identifiée dans le gène *FBN2* à l'origine de l'arachnodactylie congénitale avec contractures (CCA) ou syndrome de Beals, dont la présentation est similaire à celle observée pour le syndrome de Marfan (syndrome Marfan-like). Cette mutation, *FBN2* c.3974-26T>G localisée en position -26 dans l'intron 30 se situe à proximité de la boite de branchement s'étendant des positions -21 à -15. Il a été démontré que celle-ci induit le saut partiel de l'exon 31. La recherche de cette mutation au sein d'une même famille chez 30 individus sur 5 générations a permis de démontrer que la mutation ségrége avec la maladie pour la totalité des 18 patients avec un phénotype CCA (Maslen *et al.*, 1997; Wang *et al.*, 1995; pour revue : Lewandowska, 2013). Quelques exemples d'altérations de la région riche en pyrimidine ont été décrits dans la littérature, notamment les variations 392-8T>G et c.392-9T>G localisées dans l'intron 4 du gène *FIX*. Ces variations induisent un saut de l'exon 5 associé à une déficience du facteur FIX à l'origine de l'hémophilie B (Montejo *et al.*, 1999; pour revue : Lewandowska, 2013).

Cependant, contrairement aux mutations affectant les sites d'épissage, les mutations qui altèrent la séquence de la boite de branchement sont très rares, avec moins de 20 mutations décrites jusqu'à 2013 (pour revue : Lewandowska, 2013). De même, extrêmement peu d'altérations de la région riche en pyrimidine ont été décrites (pour revue : Lewandowska, 2013). En effet, ces altérations ne font pas, dans le cadre du diagnostic, l'objet d'une recherche mutationnelle systématique. Ce type de mutations est principalement recherché et détecté lorsque aucune altération n'est identifiée par séquençage dans la région codante, dans les sites 5' et 3' d'épissage et dans les régions les 5 'et 3' UTR (pour revue : Lewandowska, 2013). De plus, très peu de boites de branchement ou de régions riches en pyrimidines ont été identifiées expérimentalement et la dégénérescence des motifs consensus chez l'homme limitent le développement d'outils bioinformatiques permettant l'identification et la caractérisation de ce type de signaux.

- Éléments de régulation d'épissage

Les variations nucléotidiques peuvent également affecter d'autres signaux d'épissage auxiliaires : les éléments régulateurs de l'épissage. L'une des mutations de régulation d'épissage les mieux caractérisées est la variation synonyme c.840C>T du gène *SMN2* située en position +6

de l'exon 7 et à l'origine du saut de cet exon (Cartegni *et al.*, 2006; Kashima *et al.*, 2007). Le gène *SMN2* résulte de la duplication génique du gène *SMN1*, codant pour la protéine SMN (*survival of motor neurons*) impliquée dans l'amyotrophie spinale proximale (SMA, *spinal muscular atrophy*). Les gènes *SMN1* et *SMN2* ne diffèrent que par 5 nucléotides, en particulier le nucléotide en position +6 de l'exon 7 correspondant à la variation c.840C>T. Tandis que *SMN1* porteur de l'allèle « WT » (C) produit la protéine SMN pleine longueur fonctionnelle, le gène *SMN2* porteur de l'allèle « mutant » (T) produit une protéine tronquée instable. Ainsi, lorsque *SMN1* est muté à l'état homozygote ou hétérozygote composite (mutation bi-allélique), le gène *SMN2* ne compense malheureusement pas la déficience en protéine SMN, impliquée dans la biogenèse des snRNPs, induisant ainsi la SMA. Le saut d'exon 7 de *SMN2* résultant du changement nucléotidique c.840C>T peut être expliqué par deux modèles différents co-existants non exclusifs : soit une création d'un élément ESS reconnu par hnRNPA1 (Kashima and Manley, 2003), soit une destruction d'un élément ESE reconnu par SRSF1 (SF2/ASF) (Cartegni and Krainer, 2002; Cartegni *et al.*, 2006).

Bien qu'une partie du code régissant l'épissage ait été décrypté, il n'est toujours pas possible de déterminer l'effet d'une mutation sur les éléments régulateurs d'épissage uniquement à partir de la séquence génomique (pour revue : Cooper *et al.*, 2009). Ainsi, contrairement aux mutations affectant les sites d'épissage, relativement peu de mutations de régulation d'épissage associées à des maladies ont été décrites à ce jour. Pourtant, dans une enquête mutationnelle récente de la base de données HGMD, il a été estimé que 25% des variations de type faux-sens et non-sens altèrent l'épissage par création ou par destruction d'éléments exoniques régulateurs d'épissage (Sterne-Weiler *et al.*, 2011). Par ailleurs, des analyses mutationnelles ciblées sur des exons modèles ont révélées que les mutations affectant les éléments régulateurs de l'épissage seraient très probablement sous-estimées. En effet, une proportion importante des variations ponctuelles exoniques analysées dans ces exons (jusqu'à ~80%) altère l'épissage, suggérant que ces exons présentent une sensibilité particulière aux mutations qui affectent la régulation d'épissage (Soukarieh *et al.*, 2016, pour revue : López-Bigas *et al.*, 2005; Savisaar and Hurst, 2017). Il s'agit notamment de l'exon 10 de *MLH1* (10/15, 66%), l'exon 7 de *BRCA2* (11/32, 35%), l'exon 9 de *NF1* (25/35, 71%), l'exon 9 de *CFTR* (22/44, 77%), l'exon 7 de *SMN2* (34/43, 79%), l'exon 5 de *WT1* (89/139, 64%) et l'exon 9 de *FIX* (12/17, 71%), impliqués respectivement dans le syndrome de Lynch, le syndrome seins-ovaires, la neurofibromatose type I, la mucovuscidose, la SMA, les

tumeurs de Wilms et l'hémophilie B (Cartegni *et al.*, 2006; Di Giacomo *et al.*, 2013; Hernández-Imaz *et al.*, 2015; Ke *et al.*, 2018; Pagani *et al.*, 2003a; Singh *et al.*, 2007; Soukarieh *et al.*, 2016, 2016; Tajnik *et al.*, 2016). Dans leur ensemble, ces données démontrent ainsi qu'une variation génétique identifiée chez un patient est, en dépit de l'effet prédit sur la structure ou la fonction de la protéine, toujours un candidat potentiel pour être une mutation d'épissage. Par conséquent, toute variation ponctuelle, aussi bien intronique qu'exonique, qu'elle corresponde à une substitution ou une délétion/insertion, et quel que soit son impact potentiel sur la protéine (synonymes, faux-sens, non-sens) est donc susceptible d'altérer l'épissage et peut être à l'origine de maladies.

b. Altérations des facteurs trans d'épissage (trans-acting mutation)

Alors que les mutations pathogènes touchant les éléments cis affectent l'épissage d'un seul gène, les mutations touchant les composants de la machinerie d'épissage peuvent être à l'origine d'une dérégulation de l'épissage de nombreux gènes (pour revue : Ward and Cooper, 2010). Cependant, contrairement au nombre de plus en plus élevé de mutations d'épissage pathogènes touchant les éléments *cis* de l'épissage, les exemples de mutations d'épissage dans les facteurs *trans* à l'origine de maladies restent très limités (pour revue : Singh and Cooper, 2012). Par ailleurs, le faible taux de mutations dans les facteurs *trans* d'épissage suggère que les mutations affectant les composants de la machinerie d'épissage sont létales durant le développement embryonnaire (pour revue : Singh and Cooper, 2012). Cependant, quelques études ont récemment décrit des altérations dans les facteurs *trans* d'épissage, y compris dans les constituants du splicéosome et les facteurs régulateurs d'épissage, altérations qui pourraient être à l'origine de maladies tissu-spécifiques (pour revue : Daguenet *et al.*, 2015).

- Composants du splicéosome

Les altérations des composants du splicéosome peuvent être dues à des **mutations touchant des composants essentiels du splicéosome** (pour revue : Daguenet *et al.*, 2015). Des travaux de séquençage d'exome ou de génome récemment menés chez des patients atteints de désordres hématopoïétiques, lymphoïdes ou myéloïdes, ont révélé l'existence de mutations somatiques récurrentes dans les gènes codant pour des composants essentiels du splicéosome (pour revue : Daguenet *et al.*, 2015). Plus particulièrement, des mutations somatiques associées à ces maladies

ont été identifiées dans des gènes codant pour des facteurs d'épissage impliqués dans la reconnaissance du site 3' d'épissage (SF1/BBP et ZRSR2, sous-unités du facteur U2AF) et des composants du snRNP U2 (SF3A1 et SF3B1) (Papaemmanuil *et al.*, 2011; Yoshida *et al.*, 2011). D'ailleurs, le facteur SF3B1 est le plus fréquemment retrouvé muté chez ces patients et en particulier chez les patients atteints du syndrome myélodysplasique dans les formes sidéroblastiques pour lesquelles les mutations de *SF3B1* expliquent 85% des cas et corrèlent avec un pronostic plus favorable. A l'inverse, dans la leucémie lymphocytaire chronique (LLC), les mutations du gène *SF3B1*, à l'origine de 15% des LLC, sont associées à un pronostic défavorable de résistance à la fludarabine utilisée en chimiothérapie, suggérant que les effets moléculaires de ces mutations sont hautement influencés par le contexte cellulaire (Papaemmanuil *et al.*, 2011; pour revue : Daguenet *et al.*, 2015).

D'autres altérations ont également été retrouvées dans d'autres composants du splicéosome, notamment, le tri-snRNP U4.U6/U5 (pour revues : Daguenet *et al.*, 2015; Dujardin *et al.*, 2016). En effet, des mutations du gène *EFTUD2* (U5-116 kDa), codant pour un composant du snRNP U5, sont responsables de la dysostose mandibulo-faciale avec microcéphalie, un syndrome de malformations multiples, tandis que des mutations du gène *USB1*, codant pour une protéine intervenant dans la maturation du snRNA U6, sont associées à la Clericuzio-type poikiloderma (Lines *et al.*, 2012; Shchepachev *et al.*, 2012). De même, des mutations dans les facteurs PRPF (*pre-mRNA processing factor*) 31, PRPF8, PRPF6, PRPF3, PAP-1 et SNRNP200/BRR2, impliqués dans l'assemblage ou le désassemblage du complexe tri-snRNP U4/U6.U5 du splicéosome, sont à l'origine de la rétinite pigmentaire, une maladie héréditaire dégénérative conduisant à une cécité progressive (pour revue : Mordes *et al.*, 2006).

Il est intéressant de noter que l'ensemble des mutations altérant des composants du splicéosome sont associées à des pathologies tissus-spécifiques. Il est possible que l'effet tissu-spécifique résultant des altérations des composants du splicéosome soit dû au fait que la réduction de l'activité de l'épissage serait plus nuisible dans les cellules qui se divisent rapidement dans des tissus exigeant une régénération rapide telles que les cellules de la rétine ou les cellules hématopoïétiques (Neumann *et al.*, 2010; pour revue : Daguenet *et al.*, 2015). En effet, un criblage à l'échelle du génome de facteurs importants pour la division cellulaire correcte a révélé un enrichissement substantiel en composants splicéosomaux.

Les altérations des composants du splicéosome peuvent également être induites indirectement, par des mutations au niveau de certains gènes qui affectent **la biogenèse des snRNPs**, parmi lesquels les gènes *SMN1*, *TARDBP* et *FUS/TLS* (pour revue : Daguenet *et al.*, 2015). Le gène *SMN1* code pour la protéine SMN impliquée dans la biogenèse des snRNPs, qui lorsqu'elle est altérée, conduit à la SMA, une maladie à transmission récessive caractérisée par une dégénérescence des motoneurones (pour revue : Daguenet *et al.*, 2015). En effet, la protéine SMN forme un complexe fonctionnel avec les protéines Gemins 2-8 et Unrip (*unr-interacting protein*), responsable de l'assemblage des protéines Sm dans les snRNPs (So *et al.*, 2016). Ainsi, la déficience en protéine SMN induit une altération globale du répertoire de protéines snRNPs accompagnée de profondes conséquences sur l'activité de la machinerie d'épissage dans le circuit des neurones moteurs (Zhang *et al.*, 2008).

De manière intéressante, l'altération de la biogénèse des snRNP est associée à d'autres troubles moteurs et/ou neurodégénératifs, parmi lesquels la sclérose latérale amyotrophique (SLA). Cette dernière est causée par des mutations détectées dans une vingtaine de gènes avec différentes fonctions, dont les protéines de liaison à l'ARN, FUS et TDP-43 (Renton *et al.*, 2014). FUS est une protéine hnRNP-like qui interagit avec les snARNs U1 et U2 (Gerbino *et al.*, 2013; Kwiatkowski *et al.*, 2009; Vance *et al.*, 2009). Dans la SLA, les protéines FUS mutantes sont capables de se lier aux snARNs, mais elles sont séquestrées dans le cytoplasme, ce qui engendre une réduction de la quantité des snRNPs U1 et U2 disponibles dans le noyau (pour revue : Daguenet *et al.*, 2015). De plus, la protéine FUS interagissant avec la protéine SMN, les protéines FUS mutantes semblent altérer la localisation des protéines SMN et contribuent potentiellement aux altérations d'épissage associées à la SLA (Yamazaki *et al.*, 2012). Enfin, la protéine FUS interagit avec la protéine TDP-43, dont l'altération affecte également la localisation des protéines SMN et l'abondance des snARNs (Ling *et al.*, 2015).

En outre, plusieurs mutations affectant le fonctionnement du splicéosome mineur ont été caractérisées, dont celles à l'origine d'une déficience en hormone de croissance ou d'une forme de nanisme (Syndrome de Taybi-Linder ou nanisme microcéphalique ostéodysplasique primordial de type I) ou encore la SLA (Argente *et al.*, 2014; He *et al.*, 2011; Reber *et al.*, 2016). Le syndrome

de Taybi-Linder est une maladie résultant de mutations récessives du gène *MOPDI* codant pour le snRNA U4atac, composant essentiel du splicéosome mineur. A la différence des snRNA du splicéosome majeur, les snRNA du splicéosome mineur sont exprimés à partir d'un locus unique dans le génome. Ainsi, les mutations survenant au niveau des gènes codant les snRNA « U12 » déstabilisent potentiellement la fonction de ces composants. En particulier, les mutations survenant dans le gène *MOPD1,* au niveau de la structure tige-boucle présente en 5', altèrent la formation du tri-snRNP U4atac/U6atac à l'origine de cette pathologie (pour revue : Turunen *et al.*, 2013).

- Facteurs régulateurs d'épissage

L'implication d'une altération des facteurs régulateurs de l'épissage a été particulièrement bien décrite dans le cancer et les maladies neurodégénératives (pour revues : Daguenet *et al.*, 2015; Fredericks *et al.,* 2015). Même un changement modéré dans l'expression de l'un des RBPs, et donc dans sa stœchiométrie, peuvent avoir des effets significatifs sur l'épissage (pour revue : Cooper *et al.*, 2009). Ces changements peuvent être la conséquence directe d'une mutation du gène codant la RBP ou indirecte via une modification dans la régulation de leur expression, notamment par phosphorylation et/ou par leur localisation subcellulaire (pour revue : Cooper *et al.*, 2009).

De nombreuses protéines régulatrices de l'épissage, notamment des membres de la famille SRs, parmi lesquels SRSF1, SRSF2, SRSF6, de la famille des hnRNPs, notamment hnRNP A1, hnRNP I, hnRNP H ou d'autres familles de RBPs tels que TIA-1/TIAR, Sam68, HuR, NOVA, SON, RBM5 et RBM10 se trouvent dérégulées dans les cancers (pour revue : Daguenet *et al.*, 2015). Par exemple, des mutations retrouvées dans le gène codant la protéine SRSF2 contribue à la myélodysplasie. En modifiant la spécificité des RBP, ces mutations vont conduire à un changement des intéractions avec des ESR, conduisant ainsi des dérégulations de l'expression des régulateurs hématopoïétiques clés (Kim *et al.*, 2015; Komeno *et al.*, 2015; Zhang *et al.*, 2015). De même, les RBP hnRNP I/PTB et hnRNP A1/B2, surexprimées dans les glioblastomes via l'activation de l'oncogène c-myc, provoquent une modification de l'épissage alternatif du gène *PKLR* codant pour la pyruvate kinase résultant en une production d'énergie efficace dans les cellules cancéreuses par la glycolyse aérobie (David *et al.*, 2010). De plus, l'implication de plusieurs gènes codant pour des RBPs, régulant de nombreux événements d'épissage dans les neurones, a été décrite dans plusieurs maladies neurodégénératives, incluant la SLA et la

dégénérescence lobaire fronto-temporale (FUS et TDP-43), l'autisme (RBFOX1), la maladie de Huntington (SRSF6) et des troubles neurologiques paranéoplasiques (NOVA) (pour revues : Daguenet *et al.*, 2015; Fredericks *et al.*, 2015).

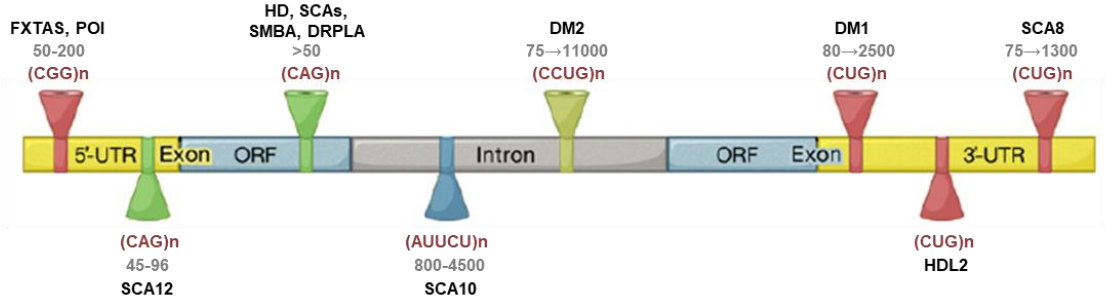
c. ARN toxique (Expansions de nucléotides répétés)

La fixation des protéines de liaison à l'ARN (RBPs) sur leurs sites spécifiques peut être altérée par l'expansion de séquences répétées dans l'ARN : c'est le mécanisme de « l'ARN toxique » (pour revues : Cooper *et al.*, 2009; Fredericks *et al.*, 2015; Mirkin, 2007). Le séquençage du génome humain a révélé que ce dernier contenait environ 3% de séquences microsatellites, courtes séquences (généralement de 1 à 4 paires de bases) répétées polymorphiques. Le nombre de ces répétitions varie généralement entre 5 et 50 chez les individus sains et reste stable au cours des générations. Cependant, le nombre de ces répétitions peut être instable et hypervariable à cause d'erreurs au cours de la réplication, de la réparation et de la recombinaison et, au-delà d'un nombre de répétitions seuil (généralement ~100-150 nucléotides), spécifique à chaque maladie à répétitions de nucléotides, ces séquences répétées deviennent pathogènes (pour revue : Pearson *et al.*, 2005). Une fois le seuil pathologique dépassé, le nombre de répétitions augmente d'une génération à la suivante. De ce fait, les générations successives sont atteintes de façon plus précoce et présentent des symptômes plus sévères de la maladie : c'est le phénomène d'anticipation (pour revue : Pearson *et al.*, 2005).

La majorité des pathologies liées à l'expansion de séquences répétées sont associées à des expansions de triplets (CNG, GAA, TTC, GCN, NGC), de quadruplets (CCTG ou CAGG), de pentanucléotides (ATTCT ou AGAAT), voire même de répétitions allant jusqu'à 12 nucléotides (C_4GC_4GCG ou $CGCG_4CG_4$) et peuvent être localisées au niveau des régions codantes des gènes (exons) ou non codantes (introns et 5' et 3'-UTR) (Figure 41 ; pour revue : Mirkin, 2007). A l'heure actuelle, une trentaine de pathologies liées à des expansions de ce type ont été clairement identifiées, en particulier des maladies dégénératives musculaires et neuronales (la chorée de Huntington, dystrophies myotoniques 1 et 2, par exemple). Cependant, la liste est loin d'être exhaustive (pour revue : Mirkin, 2007). En fonction de leur localisation au sein du gène, ces répétitions peuvent entraîner l'apparition de la maladie selon 3 mécanismes qui ne sont pas mutuellement exclusifs : (i) une perte de fonction du gène lorsque l'expansion touche une région

non-codante du gène, (ii) un gain de fonction protéique toxique lorsque l'expansion touche une région codante, (iii) un gain de fonction de l'ARN, lorsque les expansions touchent une région non-codante (régions 5' et 3' UTR ou introniques) (pour revue : Fredericks *et al.*, 2015). Lorsque ces répétitions induisent un gain de fonction de l'ARN, celles-ci peuvent entraîner des dérégulations de l'épissage du pré-ARNm. En effet, ces expansions créent des sites de fixation en tandem qui recrutent et séquestrent, comme une éponge, certains RBPs au niveau de ces transcrits (pour revue : Fredericks *et al.*, 2015). Par conséquent, les RBPs « libres » restant dans la cellule ne sont pas suffisants pour réguler l'épissage par fixation sur leurs sites spécifiques au niveau du pré-ARNm.

Figure 41 : Exemples de pathologies liées à l'expansion de nucléotides répétés et touchant différentes régions des gènes (d'après Cooper *et al.*, 2009)**.** DM1/2, dystrophine musculaire de type 1/2 ; DRPLA, Atrophie dentato rubro pallido luysienne ; FXTAS, syndrome du tremblement-ataxie lié à l'X fragile ; HD, maladie de Huntington ; HDL2, maladie de Huntington like de type 2 ; POI, insuffisance ovarienne



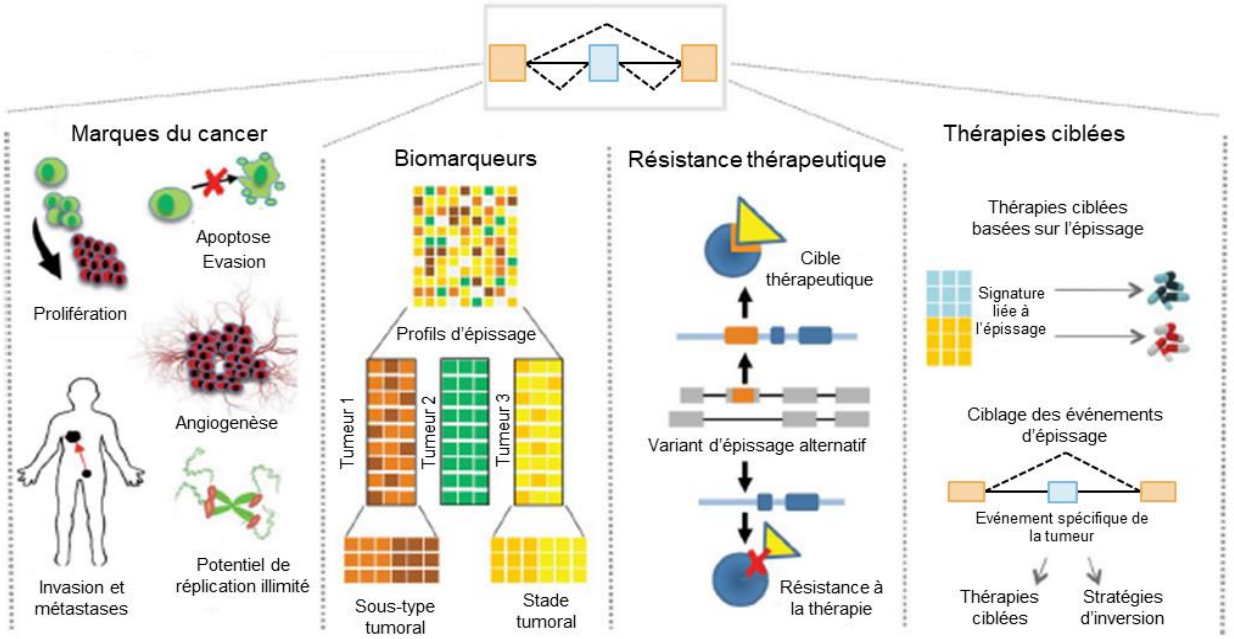
prématurée ; SBMA, atrophie musculaire spinale et bulbaire ; SCA8, ataxie spinocérébelleuse.

2) Dérégulation de l'épissage dans les cancers

Dans les cellules cancéreuses, l'épissage et plus particulièrement l'épissage alternatif, est fréquemment dérégulé à l'avantage de ces cellules. En effet, de nombreux variants d'épissage associés à la transformation tumorale ont été identifiés et sont communément retrouvés enrichis dans les tissus cancéreux, comparativement aux tissus normaux environnants (pour revues : Kim and Kim, 2012; Ladomery, 2013; Oltean and Bates, 2014; Omenn *et al.*, 2013). Des études pangénomiques ont d'ailleurs révélé l'existence de profils d'épissage spécifiques des cellules tumorales, représentant une véritable signature du cancer, ceux-ci pouvant être utilisés comme biomarqueur diagnostic (identification des sous-types tumoraux) et dans l'orientation de la

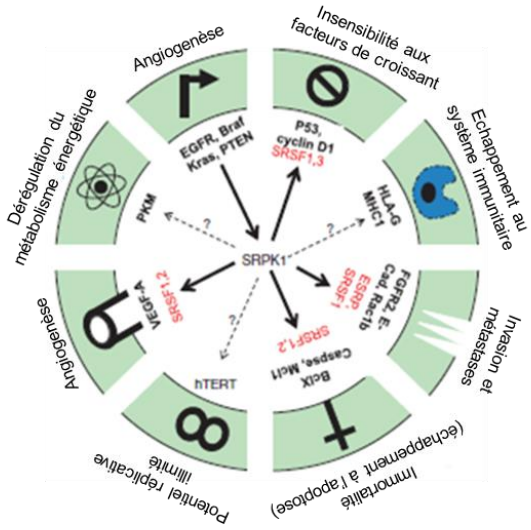
stratégie thérapeutique à adopter (Figure 42 ; pour revues : Kim and Kim, 2012; Ladomery, 2013; Oltean and Bates, 2014; Omenn *et al.*, 2013). Ces variants d'épissage peuvent résulter de mutations introniques ou exoniques présentes au niveau des signaux d'épissage survenant dans des gènes associés au cancer, tels que les proto-oncogènes et les gènes suppresseurs de tumeurs. Cependant, dans la plupart des cancers sporadiques, les gènes épissés de manière aberrante ne sont pas mutés, indiquant que les effets observés sont associés à une dérégulation de l'environnement régulant le choix du site d'épissage (Futreal *et al.*, 2004). Même si les origines exactes de la plupart de ces événements d'épissage aberrants ne sont pas connues, des facteurs d'épissage, dont les protéines SR particulièrement, sont probablement très impliqués. En effet, l'analyse d'un large panel de tumeurs, notamment des tumeurs du poumon, du colon et du sein, a permis d'identifier la protéine SRSF1 comme proto-oncogène (pour revues : Daguenet *et al.*, 2015; Wang and Cooper, 2007). La surexpression de cette protéine ou des changements dans le statut de phosphorylation, est suffisante pour induire la transformation cellulaire en modulant l'épissage alternatif des gènes suppresseurs de tumeurs et des gènes codant des protéines kinases (Figure 43), parmi lesquels S6K1 résultant en la transformation cellulaire et RON, conduisant à une augmentation de la mobilité et des métastases (Ghigna *et al.*, 2005; Karni *et al.*, 2007).

Figure 42 : Implication des altérations de l'épissage dans le cancer (d'après Singh et Eyras, 2017).



Le changement de profils d'épissage peut être critique, étant donné que de nombreux gènes associés à la progression tumorale possèdent des variants d'épissage qui sont mutuellement antagonistes (pour revues : Liu and Cheng, 2013; Oltean and Bates, 2014; Sveen *et al.*, 2016). En effet, ces gènes produisent généralement deux variants d'épissage, l'un favorisant la progression tumorale (anti-apoptotique) et l'autre l'inhibant au contraire (anti-apoptotique). Plus précisément, les variants d'épissage aberrants produits par les proto-oncogènes génèrent des protéines constitutivement actives ou des isoformes avec gain de fonction qui confèrent un avantage à ces cellules en terme de survie et de prolifération. A l'inverse, les gènes suppresseurs de tumeur avec des variants d'épissage aberrants génèrent des variants contenant des PTC ou avec décalage de lecture, potentiellement dégradés par le NMD, diminuant ses capacités tumeurs suppressives. Certains de ces transcrits, codant pour des protéines tronquées à effet dominant négatif peuvent échapper au NMD et les fonctions tumeurs suppressives de ces transcrits sont encore plus supprimées. La rupture de l'équilibre, maintenu par épissage alternatif, entre des variants d'épissage antagonistes d'un même gène peut affecter une multitude de gènes associés à la transformation et/ou la progression tumorale, c'est-à-dire à la quasi-totalité des caractéristiques du cancer (Figure 43 ; pour revues : Liu and Cheng, 2013; Oltean and Bates, 2014; Sveen *et al.*, 2016). Parmi ces gènes, celui codant le récepteur CD44 génère une vingtaine de transcrits par épissage alternatif. Certaines de ces isoformes dont CD44v confèrent un potentiel métastique aux cellules cancéreuses et sont surexprimées dans ces cellules. De manière remarquable, leur expression peut être utilisée à des fins diagnostique, pronostique et thérapeutique : l'expression spécifique de certaines isoformes, restreinte aux cellules tumorales, peut être utilisée comme biomarqueur ; une faible expression des isoformes CD44v est associée à une meilleure survie chez les patients atteints d'un cancer, en particulier de l'isoforme v6 dans les CCR ; le profil d'expression de CD44 dans les cellules cancéreuses est prédictif de la réponse au traitement anti-CD44 dans différentes tumeurs solides (pour revues : Prochazka *et al.*, 2014; Sveen *et al.*, 2016).

Figure 43 : Implication de l'altération de l'épissage alternatif des régulateurs clés dans le développement des cancers (d'après Oltean and Bates, 2014).



a. Altération des signaux d'épissage dans les cancers héréditaires

Les mutations qui altèrent les éléments *cis* d'épissage sont fréquemment impliquées dans les maladies génétiques dont les cancers héréditaires. Cela laisse supposer que les gènes de prédisposition au cancer sont particulièrement sensibles aux mutations d'épissage. En effet, une analyse des mutations d'épissage identifiées dans des gènes associés à des maladies génétiques ont permis d'identifier 86 gènes enrichis en mutations d'épissage, dont une majorité de gènes de prédisposition au cancer, incluant les 3 gènes majeurs impliqués dans le syndrome de Lynch (*MLH1*, *MSH2* et *PMS2*) (Rhine *et al.*, 2018). La majorité des mutations d'épissage identifiées dans des gènes de prédisposition au cancer se situe au niveau des sites consensus d'épissage, non seulement sur les dinucléotides introniques invariants mais aussi au niveau des nucléotides introniques et exoniques moins conservés (Houdayer *et al.*, 2012; Pagenstecher *et al.*, 2006; Rhine *et al.*, 2018; Thomassen *et al.*, 2012; Tournier *et al.*, 2008; Wappenschmidt *et al.*, 2012).

Plusieurs substitutions ponctuelles qui affectent les sites d'épissage ont été décrites dans le gène de prédisposition au cancer, notamment dans les gènes MMR et BRCA impliqués dans le syndrome de Lynch et le syndrome seins-ovaires, respectivement (Houdayer *et al.*, 2012; Pagenstecher *et al.*, 2006; Thomassen *et al.*, 2012; Tournier *et al.*, 2008; Wappenschmidt *et al.*, 2012). Parmi ces

mutations, la substitution ponctuelle *MSH2* c.942+3A>T, l'une des mutations pathogènes la plus fréquente des gènes MMR (111 entrées dans la base de données UMD-MSH2), induit le saut de l'exon 5 (Mangold *et al.*, 2005). De même, la variation *MLH1* c.546-2A>G située dans l'intron 6 est associée à un saut total combiné des exons 6 et 7 dans le sang d'un patient (Tanko *et al.*, 2002). D'autres types d'évènements ont été mis en évidence dans les cancers héréditaires tels que la destruction/création de sites d'épissage alternatifs/cryptiques dans les gènes BRCA et dans d'autres gènes de prédisposition au cancer (Hoffman *et al.*, 1998; Mazoyer *et al.*, 1996; pour revue : Vaz-Drago *et al.*, 2017).

D'autres variations nucléotidiques, situées en dehors des sites d'épissage, peuvent également altérer l'épissage de l'ARN, en touchant des éléments *cis* de régulation de l'épissage. Les mutations de régulation d'épissage ont été décrites dans de nombreux gènes impliqués dans les cancers héréditaires, et tout particulièrement dans les gènes MMR et BRCA (Sanz *et al.*, 2010; Tournier *et al.*, 2008). En outre, certains exons de ces gènes ont montré une sensibilité particulière aux mutations qui affectent la régulation d'épissage. Il s'agit, de l'exon 10 de *MLH1* (10/15, 66%), de l'exon 6 de *BRCA1* (12/42, 29%), des exons 7 et 18 de *BRCA2* (11/32, 35% et 9/23, 39%, respectivement) enrichis en mutations d'épissage situées en dehors des sites et d'épissage (Di Giacomo *et al.*, 2013; Fraile-Bethencourt *et al.*, 2017; Gaildrat *et al.*, 2012; Raponi *et al.*, 2011; Soukarieh *et al.*, 2016).

b. Altération de l'épissage dans les cancers sporadiques

Contrairement aux cancers héréditaires, les variations qui altèrent les éléments *cis* d'épissage sont assez peu fréquemment décrites dans les cancers sporadiques. En effet, les analyses des mutations d'épissage effectuées dans un contexte sporadique restent très limitées et moins prévalentes que celles réalisées dans un contexte de maladies Mendéliennes (Dorman *et al.*, 2014). Néanmoins, les analyses de Dorman et ses collaborateurs suggèrent qu'environ 6% des substitutions ponctuelles somatiques identifiées dans le cancer du sein correspondent à des mutations d'épissage (Dorman *et al.*, 2014). Les rares mutations d'épissage somatiques identifiées dans les tumeurs correspondent généralement à (i) des variations ponctuelles localisées au niveau des jonctions exons/introns dans les gènes suppresseurs de tumeurs parmi lesquels *TP53*, *ARID1A*, *PTEN*, *CHD1*, *MLL2* et *PITCH1*, et (ii) des variations synonymes ou

silencieuses sur le plan traductionnel qui altèrent l'épissage d'oncogènes tels que *ITK*, *ALK*, *IDH1* et *BCK6* (Supek *et al.*, 2014; pour revue : Singh and Eyras, 2017). En effet, les mutations synonymes semblent spécifiquement enrichies dans les oncogènes et non dans les gènes surpresseurs de tumeurs à l'exception de *TP53* pour lequel une fraction importante des mutations synonymes altère l'épissage (Supek *et al.*, 2014). Ce type de mutations représenteraient 6-8% de toutes les mutations *drivers* affectant les oncogènes, la moitié d'entre elles étant à l'origine d'une altération de l'épissage (Supek *et al.*, 2014).

Parmi les gènes les plus étudiés au niveau somatique, on retrouve l'oncogène *MET*, codant pour une récepteur à activité tyrosine kinase (RTK), et plus particulièrement les mutations affectant l'épissage de l'exon 14. En effet, celles-ci sont communément trouvées dans les cancers du poumon, notamment dans 3-4% des adénocarcinomes pulmonaires, représentant les mutations les plus communes après celles de l'EGFR (*epidermal growth factor receptor*) (pour revue : Pilotto *et al.*, 2017). L'exon 14 de *MET* code pour le domaine juxta-membranaire du RTK, contenant le résidu Y1003 qui sert de site de liaison à l'ubiquitine ligase CBL. (Kong-Beltran *et al.*, 2006; pour revue : Pilotto *et al.*, 2017). Le saut de l'exon 14 conduit à une diminution de l'ubiquitination et de la dégradation, augmentant la stabilité du RTK et l'activation des cibles en aval du RTK au sein la voie de signalisation de l'HGF (*hepatocyte growth factor*), parmi lesquelles les kinases PI3K, mTOR et MAPK et le facteur de transcription mTOR (pour revue : Pilotto *et al.*, 2017). De nombreuses mutations induisant le saut de l'exon 14 ont été retrouvées des cellules tumorales pulmonaires parmi lesquelles *MET* c.2942-27_2942-6del, c.3062_3082+7del et c.3082+1G>T (Kong-Beltran *et al.*, 2006). La recherche de telles mutations est indispensables pour la prise en charge des patients atteints d'un cancer du poumon car elles pourraient conditionner l'accès à certains inhibiteurs de MET (crizotinib, cabozantinib, capmatinib) En effet, ces inhibiteurs pourraient être proposés uniquement aux patients porteurs de mutations altérant l'épissage de l'exon 14, mutations qui confèreraient une sensibilité aux inhibiteurs de MET (pour revue : Pilotto *et al.*, 2017).

3) Approches thérapeutiques de modulation de l'épissage

Avec le développement des techniques de séquençage à haut-débit, le nombre de mutations identifiées chez les patients atteints de maladies Mendéliennes a considérablement augmenté, et en particulier le nombre de mutations affectant l'épissage, représentant aujourd'hui jusqu'à 50% de toutes les mutations pathogènes touchant certains gènes (Ars *et al.*, 2000; Teraoka *et al.*, 1999; pour revue : Baralle *et al.*, 2009). Ainsi, ces 10 dernières années, de nombreuses thérapies ciblant le processus d'épissage ont été développées afin de corriger les effets de certaines mutations au niveau du pré-ARNm, contournant ainsi le besoin de corriger ou remplacer l'ADN porteur de la mutation (thérapie génique) ou les cellules malades (cellules souches). Les thérapies modulant l'épissage de l'ARN sont considérées comme des approches thérapeutiques prometteuses et puissantes en raison de la large gamme de mutations qui peut être corrigée, et ce, indépendamment de la fonction du gène, de la facilité d'administration et du succès de ces approches dans le traitement de certaines maladies, certaines molécules étant actuellement en cours d'essais cliniques (Tableau 7 ; pour revues : Hammond and Wood, 2011; Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017).

a. Oligonucléotides anti-sens

L'une des premières approches utilisée pour moduler l'épissage est l'utilisation d'oligonucléotides anti-sens (ASO ou AON, *antisense oligonucleotide*), molécule guidée vers le pré-ARNm pour modifier l'épissage, en bloquant la production d'une isoforme protéique toxique ou en restaurant la production d'une protéine non fonctionnelle (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017). Les ASOs correspondent à de courtes séquences nucléiques (15-25 nucléotides) complémentaires d'une région cible spécifique présente sur le pré-ARNm et fonctionnent en formant des appariements de type Watson-Crick avec la séquence d'ARN ciblée, bloquant de manière stérique l'accès aux facteurs d'épissage. Ces molécules permettant la modulation de l'épissage sans toutefois favoriser la dégradation des transcrits cibles, elles sont également appelées « oligonucléotides de commutation d'épissage » (SSO, *splice-switching oligonucleotides*) (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017).

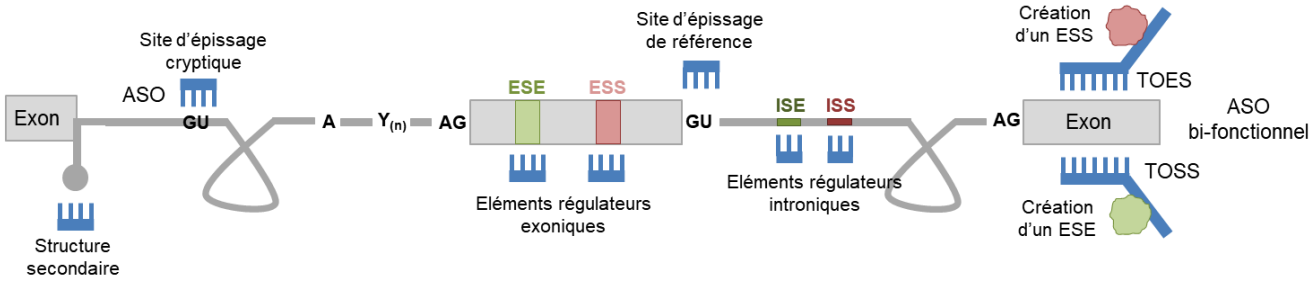
Plusieurs éléments intervenant dans la régulation du mécanisme d'épissage peuvent être cibles des ASO, en particulier (i) les sites d'épissage, naturels ou cryptiques, redirigeant le

Tableau 7 : Exemples de stratégie de modifications de l'épissage en développement dans certaines maladies génétiques (d'après Hammond and Wood, 2011).

Syndrome	Gène cible	Thérapie génique	Stade
Ataxie télangiectasie	*ATM*	AON pour la correction de l'épissage	Lignées cellulaires homozygotes
Athérosclérose	*APOB*	AON induisant le saut d'un exon	Lignée cellulaire humaine de carcinome hépatocellulaire
β-Thalassémie	β-globine	AON bloquant la reconnaissance d'un site d'épissage cryptique	Cellules mononucléées de patients modèle murin
Anomalie congénitale de la glycosylation	*PMM2*	AON bloquant la reconnaissance d'un site d'épissage cryptique	Fibroblastes de patients
Dystrophie musculaire de Duchenne	*DMD*	AON induisant le saut de l'exon 51 snRNA U7 bi-fonctionnel induisant le saut de l'exon 51 par recrutement de hnRNPA1 snRNA U1 modifié induisant le saut d'un exon	Essai clinique de phase I/IIa Myoblastes de patients Modèle murin
Dysferlinopathie	*DYSF*	AON induisant le saut d'un exon pour restaurer la phase snRNA U7 induisant le saut d'un exon par un ASO	Myoblastes de patients Myoblastes de patients
Epidermolyse bulleuse dystrophique	*COL7A1*	AON induisant le saut d'un exon contenant un PTC	Modèle murin transplanté avec de la peau humaine
Démence fronto-temporale liée au chromosome 17	*MAPT*	Trans-épissage AON bloquant la reconnaissance des sites 5' et 3' d'épissage de l'exon 10	Lignée cellulaire Fibroblastes de patients
Progéria (syndrome de Hutchinson-Gilford)	*LMNA*	AON bloquant la reconnaissance d'un site d'épissage cryptique AON induisant le saut d'un exon	Fibroblastes de patients Lignées cellulaires
Acidémie méthylmalonique	*MUT*	AON bloquant la reconnaissance d'un site d'épissage cryptique à l'origine de l'inclusion d'un pseudoexon	Fibroblastes de patients
Maladie de Pick-Niemann	*NPC1*	AON bloquant la reconnaissance d'un site d'épissage cryptique à l'origine de l'inclusion d'un pseudoexon	Fibroblastes de patients
Neurofibromatose	*NF1*	AON bloquant la reconnaissance d'un site 5' d'épissage cryptique	Lymphocytes de patients
Albinisme oculaire récessif lié à l'X	*GPR143*	AON bloquant la reconnaissance d'un élément activateur de l'épissage *de novo*	Mélanocytes de patients
Acidémie propionique	*PCCA, PCCB*	AON bloquant la reconnaissance d'un site d'épissage cryptique à l'origine de l'inclusion d'un pseudoexon	Fibroblastes de patients
Atrophie spinale musculaire	*SMN2*	AON induisant l'inclusion de l'exon 7 via le blocage d'un ISS snRNA U7 bi-fonctionnel induisant l'inclusion de l'exon 7 par recrutement de protéines SR snRNA U7 Trans-épissage SM	Modèle murin Modèle murin Fibroblastes de patients Modèle murin Essais cliniques

splicéosome vers un site d'épissage adjacent, (ii) les éléments activateurs ou inhibiteurs de l'épissage, empêchant la liaison des facteurs de régulation d'épissage afin de favoriser ou inhiber l'utilisation des sites d'épissage et (iii) des structures secondaires en particulier de type tige-boucle afin de renforcer ou inhiber la formation de ces structures (Figure 44 ; pour revues : Singh and Cooper, 2012; Spitali and Aartsma-Rus, 2012). En ciblant ces signaux, les molécules d'ASO peuvent provoquer ou corriger une variété d'anomalies en induisant notamment (i) le saut d'un exon porteur d'un PTC ou à l'origine d'un décalage du cadre de lecture comme les exons 45, 51 ou 53 du gène *DMD* impliqué dans la dystrophie musculaire de Duchenne (DMD), (ii) l'inclusion d'un exon normalement exclu par la machinerie d'épissage comme l'exon 7 du gène *SMN2* à l'origine de la SMA, (iii) l'inhibition de l'utilisation d'un site cryptique comme le site donneur cryptique de l'intron 11 du gène *LMNA* responsable du syndrome de Hutchinson-Gilford (Progéria) et (iv) l'inhibition de la formation, par des expansions nucléotidiques, de structure secondaire en épingle à cheveux comme par exemple les répétions de type CUG dans le gène *DMPK* impliqué dans la dystrophie myotonique (pour revues : Singh and Cooper, 2012; Spitali and Aartsma-Rus, 2012).

<u>Figure 44 :</u> Mécanisme de correction de l'épissage de l'ARNm à l'aide d'oligonucléotides anti-sens (adapté de Havens *et al.*, 2013). ASO, *antisense oligonucleotide* ; TOES, *targeted oligonucleotide enhancer of splicing* ; TOSS, *targeted oligonucleotide silencer of splicing.*



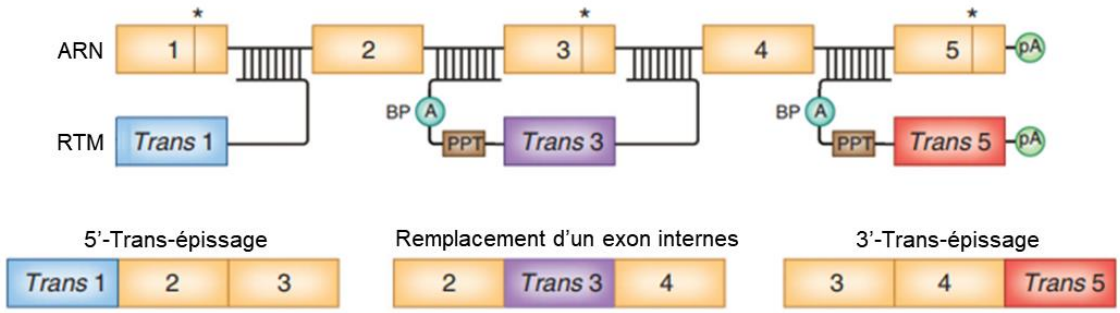
Certaines de ces ASO, approuvés en 2016 par la FDA (*<u>F</u>ood and <u>D</u>rug <u>A</u>dministration*), ont montré leur efficacité dans le traitement de certaines maladies, notamment l'Eteplirsen (Exondys 51) dans la DMD (*DMD*) et le nusinersen (Spinraza) dans la SMA (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wan and Dreyfuss, 2017). En effet, les ASO présentent de nombreux

avantages qui expliquent leur utilisation en clinique Tout d'abord, ces molécules présentent une très grande spécificité, permettant de cibler des isoformes d'ARN distinctes ou les transcrits générés à partir de l'allèle porteur de la mutation uniquement, tout en étant non invasives (elles ne modifient ou n'altèrent pas le génome). De plus, les ASO sont des molécules très efficaces, très peu toxiques et spontanément internalisées par les cellules *in vivo*. Elles sont également très stables, la demi-vie actuelle s'étendant entre 10-15 jours actuellement. Cette stabilité peut être allongée grâce à l'utilisation de modifications chimiques particulières qui vont protéger les ASO de la dégradation par les nucléases de type RNases. D'ailleurs la pharmacocinétique et la pharmacodynamique de ces molécules peuvent être modifiées afin d'améliorer cette demi-vie et des adresser vers le bon tissu. D'autres modifications peuvent être apportées aux ASO lorsque l'appariement de bases ne serait pas suffisant pour modifier l'épissage de manière efficace. En effet, des séquences nucléotidiques non appariées avec l'ARNm servant de plateforme pour le recrutement de facteurs activateurs (TOES, *targeted oligonucleotide enhancer of splicing*) ou inhibiteurs de l'épissage (TOSS, *targeted oligonucleotide silencer of splicing*), en fonction qu'elles contiennent des sites de liaison pour les protéines SR ou hnRNP peuvent être ajoutées aux ASO (Figure 44). Ces séquences vont ainsi permettre aux ASO d'assurer une dualité de fonction (bi-fonction) en bloquant d'une part le recrutement des facteurs d'épissage via l'ASO lui-même et en favorisant, d'autre part, le recrutement de protéines régulatrices de l'épissage (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017).

b. Trans-épissage

L'épissage peut être également modulé en reprogrammant un ARNm via le remplacement, dans les cellules, de la partie du transcrit muté qui doit être corrigée. Il s'agit du trans-épissage à médiation splicéosomale (SMaRT, *spliceosomal-mediated RNA trans-splicing*) (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wally *et al.*, 2012). Cette approche permet alors la conversion de l'allèle mutant en allèle sauvage au niveau de l'ARN, ce qui permet à la fois de neutraliser l'expression de l'allèle dominant-négatif et d'augmenter les niveaux d'expression de la protéine sauvage. Pour ce faire, une molécule de pré-trans-épissage (PTM, *pre-trans-splicing molecule* ou RTM, *RNA trans-splicing molecule*), introduite dans la cellule, délivre une portion de la région codante sauvage exogène qui va se substituer, par trans-épissage, à la portion mutée

d'intérêt présente sur l'ARN endogène (Figure 45 ; pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wally *et al.*, 2012). Ces molécules sont souvent assimilées à des ASOs car elles contiennent une séquence cible de reconnaissance, également appelé domaine de liaison, capable de s'hybrider spécifiquement au pré-ARNm endogène d'intérêt. Cependant, ces molécules possèdent également des caractéristiques qui leur sont propres : (i) une copie exogène de la portion de la séquence d'ARN qui doit être remplacée, (ii) des sites d'épissage fonctionnels et (iii) des signaux d'épissage auxiliaires (point de branchement et région riche en pyrimidine) qui redirige la machinerie d'épissage depuis l'ARN endogène vers la PTM (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wally *et al.*, 2012).



<u>Figure 45 :</u> Mécanisme de trans-épissage de l'ARN utilisé à visée thérapeutique pour remplacer une portion du gène (d'après Wally *et al.*, 2012).

Le trans-épissage est un mécanisme médié par la cellule elle-même et en particulier par le splicéosome. En effet, bien que le trans-épissage soit un événement rare, il survient naturellement chez l'homme, notamment pour le récepteur aux œstrogènes humain (Flouriot *et al.*, 2002; Wally *et al.*, 2012). Seule la distribution de la PTM est nécessaire, ce qui pose les mêmes limitations que celles rencontrées lors de n'importe quelle introduction de matériel génétique dans les cellules. Néanmoins, la séquence nucléotidique à introduire est nettement plus courte que celles généralement utilisées pour le transfert de gêne (gène entier, parfois trop grand pour être correctement empaqueté dans un vecteur), permettant d'adresser plus efficacement ces molécules (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wally *et al.*, 2012). De plus, le trans-épissage permet de garantir que le profil d'expression du gène cible soit conservé quantitativement, spatialement et temporellement. En effet, le PTM ne peut interagir qu'avec une molécule de pré-

ARNm préexistant, produit par la cellule elle-même et dont l'expression reste sous contrôle du promoteur endogène (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wally *et al.*, 2012).

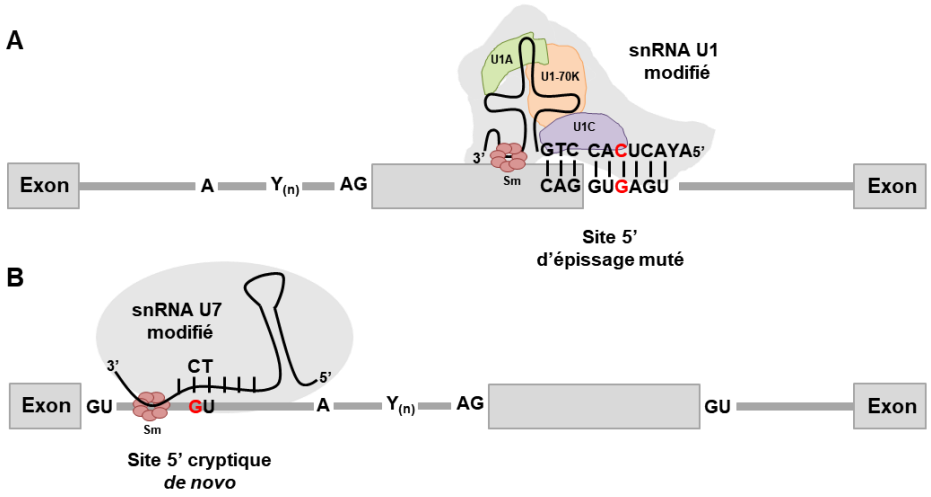
3 types d'approches SMaRT ont été développées pour restaurer l'effet des mutations affectant les sites d'épissage : le 5'-trans-épissage, le 3'-trans-épissage et le remplacement d'exons internes (IER, *internal exon replacement*) ciblant respectivement le site 5', 3' ou un ou plusieurs exons internes d'un pré-ARNm muté (pour revues : Havens *et al.*, 2013; Suñé-Pou *et al.*, 2017; Wally *et al.*, 2012). Celles-ci ont été utilisées pour la correction *in vitro* et/ou *in vivo* dans des modèles murins, d'un nombre restreint de gènes parmi lesquels *SMN2* (SMA), *F8* (hémophilie A), *MAPT* (démences frentotemporales et tauopathies) ou *CFTR* (mucoviscidose) (Wally *et al.*, 2012).

c. snRNA modifiés

D'autres approches, basées sur l'utilisation de snRNA modifiés, en particulier les snRNA U1 et U7, ont été développées pour reverser l'effet, sur l'épissage, de certaines mutations (Figure 46). Le snRNA U1 est un composant du splicéosome qui participe à la reconnaissance du site 5' d'épissage. En effet, ce dernier interagit physiquement avec le site 5' d'épissage via un appariement de base spécifique et cette complémentarité influence la sélection du site 5' d'épissage (pour revues : Wahl *et al.*, 2009; Will and Lührmann, 2011). Une mutation sur le site 5' d'épissage peut ainsi compromette la liaison du snRNA U1, empêchant ainsi l'assemblage du splicéosome et un épissage correct de l'intron contenant la mutation (pour revues : Wahl *et al.*, 2009; Will and Lührmann, 2011). Néanmoins, les mutations survenant sur le site 5' d'épissage peut être corrigées, en restaurant un épissage normal, grâce à l'utilisation de snRNAs U1 artificiels (pour revues : Hammond and Wood, 2011; Havens *et al.*, 2013). Ces molécules correspondent à une version modifiée du snRNA U1, comportant des changements au niveau de la séquence, de manière à restaurer la complémentarité de base avec le site 5' d'épissage muté afin que ce snRNA modifié se lie, de manière efficace, sur le site d'épissage muté du pré-ARNm (pour revues : Hammond and Wood, 2011; Havens *et al.*, 2013). Cette approche a été utilisée avec succès *in vitro* pour le traitement d'anomalies d'épissage liée à des mutations au niveau du site 5' d'épissage pour de nombreuses maladies, incluant la mucoviscidose (*CFTR*, stade minigène), l'hémophilie A (*FIX*, stade minigène), l'anémie de fanconi (*FANCC*, stade cellules de patients), le syndrome de Bardet-Biedl (*BBS1*, stade cellules de patients) (pour revues : Hammond and Wood, 2011; Havens *et al.*,

2013). Cependant, jusqu'à maintenant, les snRNAs U1 modifiés n'ont jamais permis de corriger, de manière efficace, des altérations de l'épissage dues à des mutations affectant le premier ou le deuxième nucléotide de l'intron, sûrement parce que ces nucléotides sont spécifiquement requis pour la réaction catalytique d'épissage (pour revues : Hammond and Wood, 2011; Havens *et al.*, 2013).

Figure 46 : Mécanisme de correction de l'épissage de l'ARNm à l'aide de snRNA modifié (adapté de Havens *et al.*, 2013). (A) Compensation d'une mutation sur le site 5' d'épissage par l'utilisation d'un snRNA U1 modifié. (B) Restauration d'un épissage physiologique à l'aide du snRNA U7 modifié.



Une stratégie à visée thérapeutique basée sur des snRNA a également été développé. Il s'agit de modifications de la particule non-splicéosomale snRNA U7 (Figure 46 ; Gorman *et al.*, 1998). Contrairement au snRNA U1, le snRNA U7 n'est pas intrinsèquement impliqué dans l'épissage des pré-ARNm mais dans le maturation des ARNm d'histones (pour revue : Schümperli and Pillai, 2004). Cependant, en changeant la séquence du snRNA U7 sur laquelle se lient les protéines Sm, protéines de liaison à l'ARN, le snRNA U7 peut alors être converti en un facteur modulateur de l'épissage artificiel qui induit soit l'inclusion, soit l'exclusion d'un exon, dépendant de la séquence ciblée sur l'ARNm et de la présence d'éventuelles modifications supplémentaires (pour revues : Hammond and Wood, 2011; Schümperli and Pillai, 2004). En effet, les snRNA U7 modulent l'épissage indirectement, via des séquences de type ASO incluses dans les snRNA U7 qu'ils

délivrent, de façon à permettre la reconnaissance spécifique de la séquence d'intérêt au sein du pré-ARNm et/ou le blocage de certains signaux d'épissage (sites d'épissage, points de branchement et éléments régulateurs) (pour revue : Schümperli and Pillai, 2004). Toutefois, étant incorporé dans une particule snRNP, l'activité des ASO est améliorée. En effet, l'ASO est protégé contre la dégradation, s'accumule efficacement dans le noyau où se produit l'épissage et s'incorporent plus facilement à la machinerie d'épissage, et en particulier au splicéosome (Marquis *et al.*, 2007). Même lorsqu'ils sont exprimés de manière permanente, ces snRNA U7 n'interfèrent pas avec le traitement de l'ARN des histones ni tout autres processus vitaux. Ils ne provoquent pas non plus, à priori, de réactions immunologiques, et aucun effet toxique n'a été observé, que ce soit dans des cultures cellulaires ou chez des souris transgéniques (Marquis *et al.*, 2007; pour revue : Schümperli and Pillai, 2004). Cette approche a d'ailleurs été efficacement utilisée dans la correction des mutations d'épissage survenant dans les gènes *DMD* et *SMN2*, impliqués dans la DMD et la SMA, respectivement (Geib and Hertel, 2009; Goyenvalle *et al.*, 2004).

d. Les petites molécules

Ces vingt dernières années, des petites molécules modulatrices de l'épissage (SPLM, *splice modulators*) ont été décrites (pour revue : Havens *et al.*, 2013). Ces molécules naturelles, souvent identifiées par criblage haut-débit, modulent l'épissage directement en modifiant l'activité des facteurs d'épissage ou indirectement par des mécanismes souvent méconnus (pour revue : Havens *et al.*, 2013). De nombreuses classes de petites molécules ont été décrites en fonction de leur mode d'action : (i) les agents pharmacologiques dirigés contre certains composants du splicéosome bloquant ainsi l'assemblage du splicéosome, notamment la spliceostatine, pladienolide ou meayamycine dirigées contre SF3B1, composant essentiel du snRNP U2, (ii) des composés pharmacologiques modifiant l'activité des protéines régulatrices de l'épissage en particulier les protéines SR via l'inhibition des SR protéines kinases (CLK ou DYRK) responsables de la phosphorylation des protéines SR, indispensable à leur localisation et leur activité, (iii) d'autres composés (ataluren) inhibant la surveillance et la dégradation, par le NMD, des transcrits contenant des PTC via la translecture des codons stop et (iv) des petites molécules aux effets moins spécifiques sur l'épissage et dont le mode d'action n'est pas bien caractérisé telles que le

clotrimazole, l'indole, les dérivés de la diospyrine, ou encore les camptothécines, le butyrate ou le valproate de sodium ou le resvératrol (Lin, 2017).

L'inconvénient majeur de telles molécules est leur manque de spécificité pouvant conduire à des effets *off-target*, en particulier lorsque leur mécanisme d'action est n'est pas caractérisé (pour revue : Havens *et al.*, 2013). Pourtant, l'utilisation de beaucoup de ces molécules a déjà été étudiée voire approuvée en clinique dans le traitement de certaines maladies (pour revue : Havens *et al.*, 2013). Tout d'abord, l'utilisation de bon nombre de ces molécules est étudiée dans le traitement des cancers. En effet, dans les cellules cancéreuses, l'épissage alternatif est souvent dérégulé, en particulier au niveau des gènes impliqués dans l'apoptose et la progression du cycle cellulaire. Il semblerait ainsi que les cellules cancéreuses soient ainsi particulièrement sensibles aux effets moléculaires des petites molécules modulatrices de l'épissage comparativement aux cellules normales (pour revue : Daguenet *et al.*, 2015). Par exemple, l'effet thérapeutique de la splicéostatine A a été montré dans les cellules de mélanomes devenues résistantes au vemurafenib via notamment l'altération de l'épissage de BRAF (saut des exons 4 à 8) (Salton *et al.*, 2015). D'autres molécules sont spécifiquement utilisées dans le traitement de certaines maladies génétiques, parmi lesquelles la DMD et la SMA. En effet, l'ataluren (Translarna™, PTC124), agent pharmacologique induisant la translecture des codons stops prématurés, en particulier ceux localisées dans le gène *DMD*, bénéficie depuis le 31 juillet 2014 (renouvelée en 2017) d'une autorisation de mise sur le marché (AMM) alors qu'après un essai clinique de phase III (NCT02139306) cette molécule s'est avérée inefficace dans le traitement de la mucoviscidose (Ryan, 2014). De même, de nombreuses SPLM augmentant les quantités de protéines SMN via l'inclusion de l'exon 7 du gène *SMN2* ont été identifiées, notamment un dérivé de la tétracycline (PTK-SMA1), sans que les mécanismes aient été toutefois élucidés (Hastings *et al.*, 2009).

CHAPITRE VI : METHODES D'ANALYSES POUR LA DETECTION DES ANOMALIES D'EPISSAGE

Idéalement, les anomalies d'épissage induites par des variations nucléotidiques doivent être mises en évidence par l'analyse comparative des profils d'épissage des transcrits exprimés dans les tissus du patient porteur de la variation avec ceux obtenus chez des individus témoins. L'épissage de l'ARN étant un mécanisme tissu-spécifique, il a été montré qu'une même mutation pouvait avoir des effets différents sur l'épissage selon les tissus analysés (pour revues : Baralle and Baralle, 2005; Baralle and Buratti, 2017). Malheureusement, les tissus relevant, en particulier le côlon et l'endomètre dans le cas du syndrome de Lynch et les tissus mammaires et ovariens dans le cas du syndrome seins-ovaires, ne sont presque jamais disponibles. Pour autant, le nombre de mutations à l'origine d'altérations de d'épissage identifiées dans des gènes directement impliqués dans des maladies génétiques, et notamment le syndrome de Lynch et le syndrome seins-ovaires, ne cesse d'augmenter (Fraile-Bethencourt *et al.*, 2017; Leman *et al.*, 2018; Rhine *et al.*, 2018). Les laboratoires de recherche et de diagnostic moléculaire ont développé et optimisé ces dernières années des stratégies combinant approches expérimentales et approches bio-informatiques visant à améliorer l'identification des mutations d'épissage et par conséquent le diagnostic moléculaires des maladies d'origine génétique.

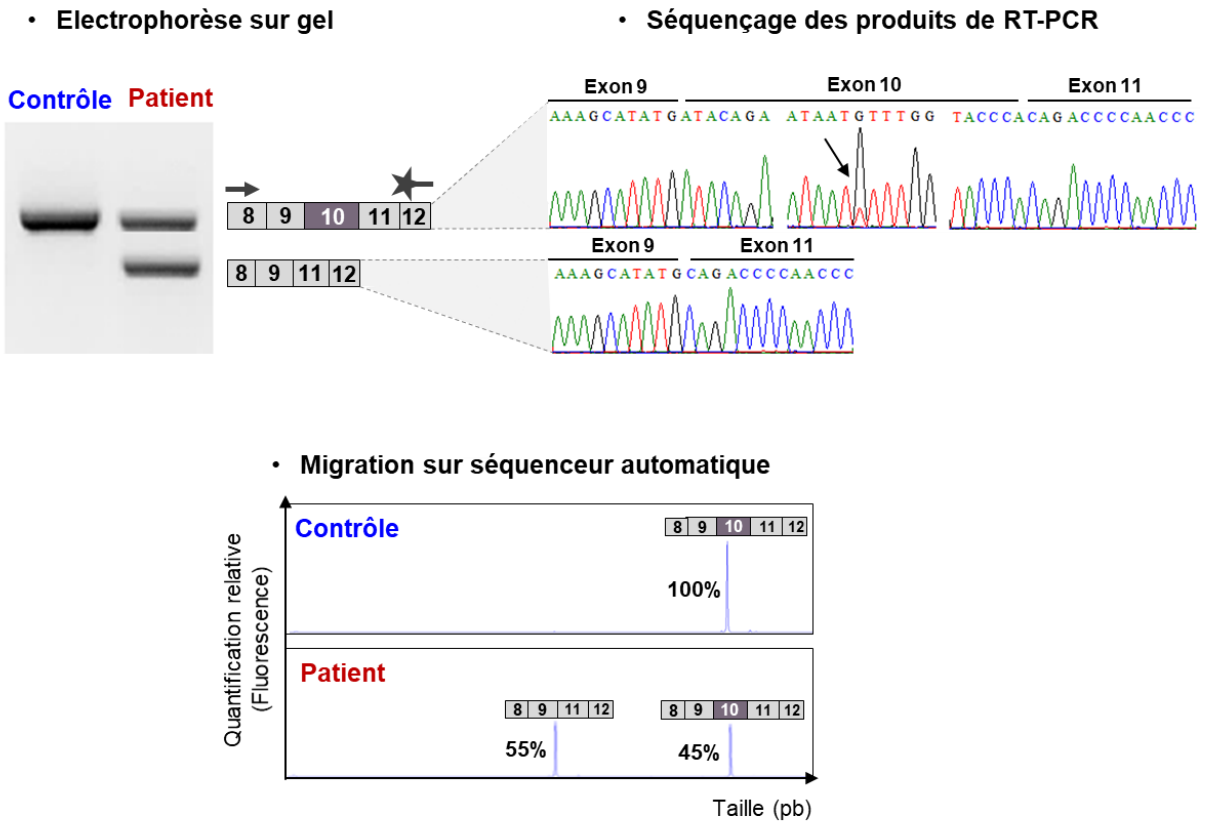
1) Analyses expérimentales basées sur l'analyse de l'ARN de patients

a. RT-PCR

L'étude des profils d'épissage des transcrits d'intérêt est réalisée, par les laboratoires de diagnostic, à partir d'échantillons de tissus relevants, idéalement lorsque ceux-ci sont disponibles, ou alternativement à partir de leucocytes issus du sang des patients ou de lignées lymphoblastoïdes, lorsque les gènes étudiés, comme les gènes MMR et BRCA, sont exprimés dans ce matériel biologique. Cette approche consiste à comparer, par RT-PCR à l'aide d'amorces positionnées dans des exons en 5' et 3' de l'exon cible, le profil d'épissage de la région d'intérêt entourant la variation

à tester à celui d'individus contrôles non porteurs de variations dans le gène analysé (Gaildrat *et al.*, 2012; pour revue : Hartmann *et al.*, 2008), après migration des produits de RT-PCR sur gel d'agarose (Gaildrat *et al.*, 2012) ou par électrophorèse sur capillaire (Figure 47 ; Romero *et al.*, 2015). Cette approche repose sur une connaissance approfondie du profil physiologique d'épissage des gènes d'intérêt. En effet, le choix des amorces est une étape cruciale pour la réaction de RT-PCR et elle est conditionné par l'existence d'épissages alternatifs dans la région d'intérêt entourant la variation à tester (Whiley *et al.*, 2014; pour revue : Baralle *et al.*, 2009). La description claire et détaillée des transcrits alternatifs des gènes d'intérêt joue un rôle très important dans le diagnostic, en particulier dans l'interprétation de l'effet sur l'épissage des variations détectées chez des patients. Dans le cas des gènes MMR et BRCA, une liste d'épissages alternatifs détectés dans ces gènes réalisée par des approches de RT-PCR, séparation sur gel d'agarose et/ou par électrophorèse capillaire a été récemment publiée (Colombo *et al.*, 2014; Fackenthal *et al.*, 2016; pour revue : Thompson *et al.*, 2015).

<u>Figure 47 :</u> Principe du test fonctionnel d'épissage, basé sur l'étude du matériel biologique du patient.



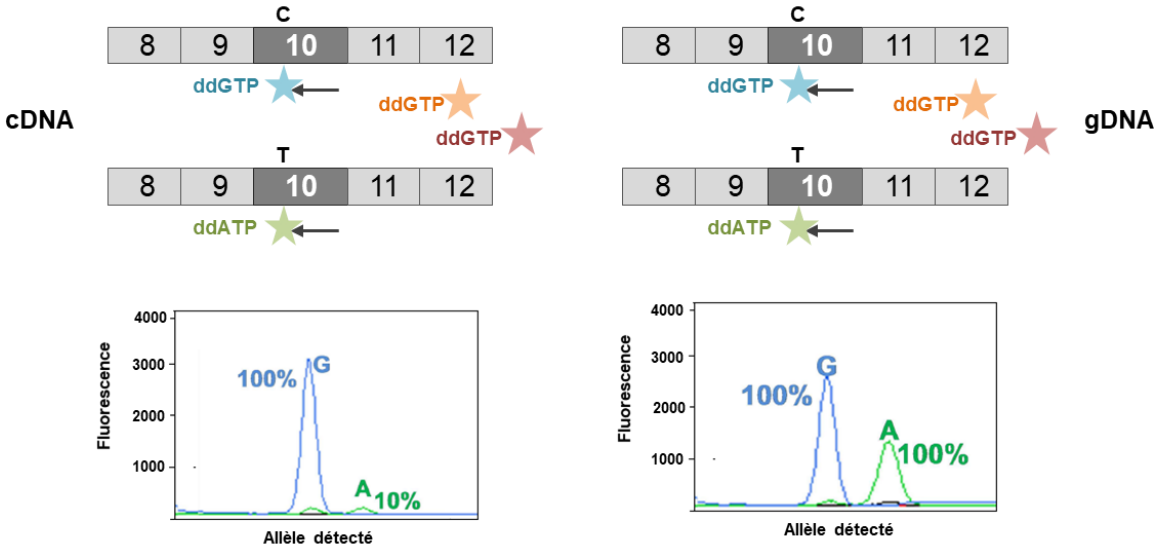
Toutefois, même si cette approche permet l'étude de l'expression endogène naturelle des gènes d'intérêt et est directement réalisée à partir de matériel biologique du patient porteur de la variation, elle présente quelques limitations, notamment au niveau de la disponibilité du matériel biologique et de la faible expression de certains gènes dans l'ARN extrait d'un tissu donné, en l'occurrence le sang périphérique (pour revues : Baralle and Buratti, 2017; Baralle *et al.*, 2009). De plus, bien que le séquençage des produits de RT-PCR permet parfois de conclure sur le caractère partiel ou total du défaut d'épissage, la quantification du déséquilibre allélique reste quant à elle peu précise (Houdayer *et al.*, 2012). Dans le cas d'une variation intronique, l'analyse de cette représentation allélique n'est le plus souvent pas possible, la variation étant non détectable au niveau de l'ADNc, à moins d'interroger un polymorphisme présent à l'état hétérozygote et à proximité de la variation. L'interprétation des données peut être également compliquée du fait du contexte bi-allélique de l'analyse mais également de la dégradation éventuelle, par le système NMD, des transcrits aberrants porteurs de PTC résultant d'un épissage alternatif ou d'un épissage aberrant provoqué par une mutation donnée (pour revues : Hartmann *et al.*, 2008; Spurdle *et al.*, 2008; Baralle and Buratti, 2017; Baralle *et al.*, 2009). Depuis quelques années, les analyses sur ARN de sang de patients se sont démocratisées suite à la mise sur le marché des tubes PAXgene (Qiagen) permettant une collecte simple et rapide d'échantillons ARN de qualité, notamment grâce à la solution qu'ils contiennent permettant de stabiliser les ARN. Cependant, les transcrits extraits à partir de ce type d'échantillons peuvent être cibles du NMD et ne sont pas facilement détectables (Houdayer *et al.*, 2012). Pour pallier à cette limite, les laboratoires de diagnostic ont également recours à l'utilisation de lignées lymphoblastoïdes, établies à partir des cellules de sang de patients. En effet, ces lignées peuvent être traitées par des inhibiteurs de la synthèse protéique (puromycine ou cycloheximide), permettant indirectement une inhibition du NMD et empêchant la dégradation de ces transcrits (Houdayer *et al.*, 2012).

b. Analyses d'expression allélique par extension d'amorces

Afin d'identifier exactement l'allèle à l'origine de l'épissage anormal et de détecter tout déséquilibre allélique éventuel, des analyses d'expression allélique (ASE, *allele specific expression*) peuvent également être réalisées sur l'ARN de patients (Caux-Moncoutier *et al.*, 2009; Tournier *et al.*, 2004; pour revue : Hartmann *et al.*, 2008). L'une des méthodes les plus utilisées est la méthode d'extension d'amorces (SNapShot®) ciblant un SNV a une position connue sur l'ADNc

et basée sur l'extension de l'extrémité 3' d'une amorce par un didésoxyribonucléotide triphosphate (ddNTP) terminateur de séquence (Figure 48). Cela implique l'hybridation de l'amorce sur l'ADNc à une base en amont ou en aval de la variation ciblée puis son extension en présence de ddNTPs fluorescents représentatifs de l'allèle sauvage et muté. Il suffit alors de normaliser la ratio allélique par rapport à celui obtenu pour l'échantillon d'ADNg du même patient (Caux-Moncoutier *et al.*, 2009; Tournier *et al.*, 2004). Cette approche a permis notamment (i) de détecter et de mesurer le déséquilibre allélique résultant de la dégradation par le NMD de l'allèle porteur d'un PTC dans les gènes BRCA (Caux-Moncoutier *et al.*, 2009) et (ii) d'évaluer la sévérité du défaut d'épissage induit par la variation MLH1 c.793C>T (Soukarieh *et al.*, 2016).

<u>Figure 48 :</u> Principe de l'analyse d'expression allélique, basée sur la méthode d'extension d'amorces (SNapShot®).



La méthode d'extension d'amorces peut être utilisée directement pour mettre en évidence le déséquilibre allélique de toute variation exonique d'intérêt. Alternativement, pour les variations introniques, non détectables au niveau de l'ADNc, le déséquilibre allélique peut être appréhendé grâce à l'interrogation de polymorphismes exoniques (SNP) dit informatifs, si présents à l'état hétérozygote et pas trop loin de la variation d'intérêt (pour revue : Hartmann *et al.*, 2008). L'utilisation de SNP exoniques communément identifiés chez des individus permet d'ailleurs

l'analyse simultanée de l'expression allélique de plusieurs individus porteurs du même SNP, réduisant le temps et le cout de l'analyse. Par exemple, au niveau des gènes BRCA, les SNPs c.2612C>T (rs799917, MAF 41% et 46% dans GnomAD et dbSNP) et c.4308T>C (rs1060915, MAF 50% et 34% dans GnomAD et dbSNP) dans les exons 11 et 13 de *BRCA1*, respectivement, et les SNPs c.3807T>C (rs543304, MAF 18% et 17% dans GnomAD et dbSNP) et c.7242A>G (rs1799955, MAF 23% dans GnomAD et dbSNP) dans les exons 11 et 14 de *BRCA2*, respectivement, sont fréquemment utilisés pour les études d'expression allélique dans ces gènes (Caux-Moncoutier *et al.*, 2009). Concernant le gène *MLH1*, il s'agit du SNP le plus fréquent, c.655A>G (rs1799977, MAF 23% et 13% dans GnomAD et dbSNP), situé dans l'exon 8. L'utilisation d'un SNP exonique pour appréhender le déséquilibre allélique d'une variation intronique implique idéalement que l'expression allélique du SNP utilisé soit établie chez des individus témoins au préalable, avant que le SNP ne soit utilisé pour détecter le déséquilibre allélique chez les patients (Buckland, 2004; Caux-Moncoutier *et al.*, 2009; Tournier *et al.*, 2004).

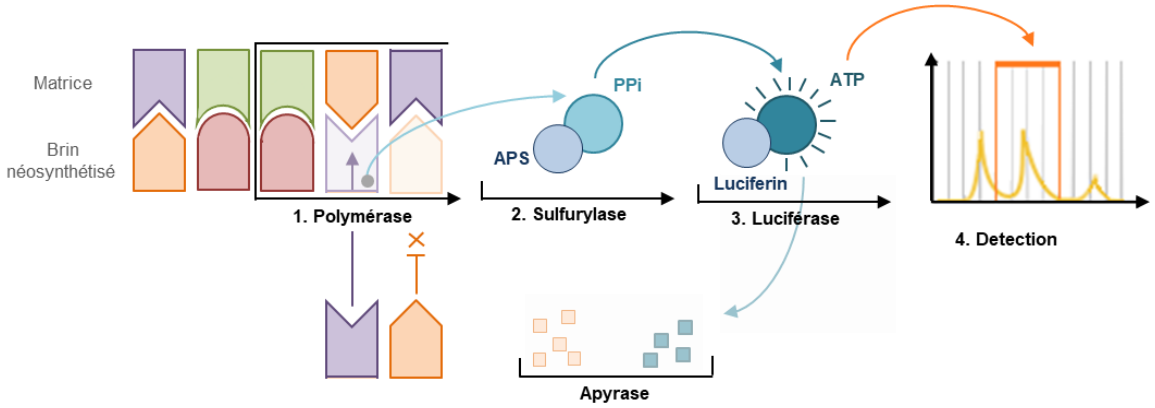
c. Analyses d'expression allélique par pyroséquençage

Un déséquilibre allélique peut également être évalué, de façon plus fine, par la méthode de pyroséquençage (Gaildrat *et al.*, 2012; Kwok *et al.*, 2010; Muller *et al.*, 2011a; Volard *et al.*, 2012). Cette méthode utilise la détection de lumière produite par l'enzyme luciférase après conversion enzymatique de pyrophosphate en ATP, le pyrophosphate résultant de l'incorporation d'un dNTP sur le brin d'ADNc (Figure 49). Le pyroséquençage peut être réalisé directement sur les variations exoniques d'intérêt (Gaildrat *et al.*, 2012) ou, de la même manière que la méthode d'extension d'amorces, peut cibler un SNP exonique hétérozygote, idéalement situé à proximité de la variation d'intérêt (Kwok *et al.*, 2010).

Le pyroséquençage a d'ailleurs été utilisé pour mettre en évidence la perte d'hétérozygotie au niveau de la tumeur de patients atteints du syndrome de Lynch, pour lesquels une perte d'expression de la protéine MLH1 avait été observée mais pas de mutation constitutionnelle causale. En exploitant le SNP le plus commun décrit dans le gène *MLH1* c.655A>G (rs1799977), situé dans l'exon 8, le pyroséquençage a permis de quantifier les taux de transcrits générés par les deux allèles chez les patients hétérozygotes (Kwok *et al.*, 2010). De même, cette méthode a permis

d'appréhender de manière quantitative la sévérité les défauts d'épissage induits par deux variations, c.520C>T et c.617C>G, situées dans l'exon 7 de *BRCA2* en (Gaildrat *et al.*, 2012).

Figure 49 : Principe de l'analyse d'expression allélique, basée sur la méthode de pyroséquençage.



d. Techniques d'analyses globales et ciblées du transcriptome : RNA-seq

Récemment, de nouvelles méthodes d'analyse globale du transcriptome ont fait leur apparition, basées notamment sur le séquençage à haut débit de l'ARNm (RNA-Seq). Il est aujourd'hui possible, notamment grâce au RNA-seq, d'étudier à large échelle les aspects quantitatifs et qualitatifs du transcriptome, c'est-à-dire d'étudier à la fois les niveaux d'expression et le type d'isoformes exprimées dans un tissus donné afin d'explorer la complexité du transcriptome (Marioni *et al.*, 2008; Wang *et al.*, 2009). En plus de permettre la comparaison des changements d'expression génique en réponse à la différenciation cellulaire, à des facteurs environnementaux ou à des conditions de maladie, le RNA-Seq peut être utilisé pour identifier avec précision de nouvelles isoformes, évaluer l'abondance relative des transcrits et détecter une autre utilisation des sites d'épissage et d'exon dans les tissus ou les cellules (Marioni *et al.*, 2008; Wang *et al.*, 2009). Cependant, à l'échelle du transcriptome entier, une minorité de gènes hautement exprimés constituent la majorité des molécules d'ARN d'une cellule. Par conséquent, le RNA-Seq global n'obtient qu'une couverture éparse des transcrits faiblement exprimés, ne permettant pas une analyse précise de l'expression de ces gènes. Pour pallier à cette limitation, il a été développé une technique alternative de RNA-Seq ciblé qui consiste à concentrer le séquençage sur des gènes d'intérêt, offrant ainsi un enrichissement considérable de la couverture et de la profondeur de

lecture ces gènes. Cela permet une analyse de l'expression des gènes plus sensible des transcrits même faiblement exprimés (Mercer *et al.*, 2014).

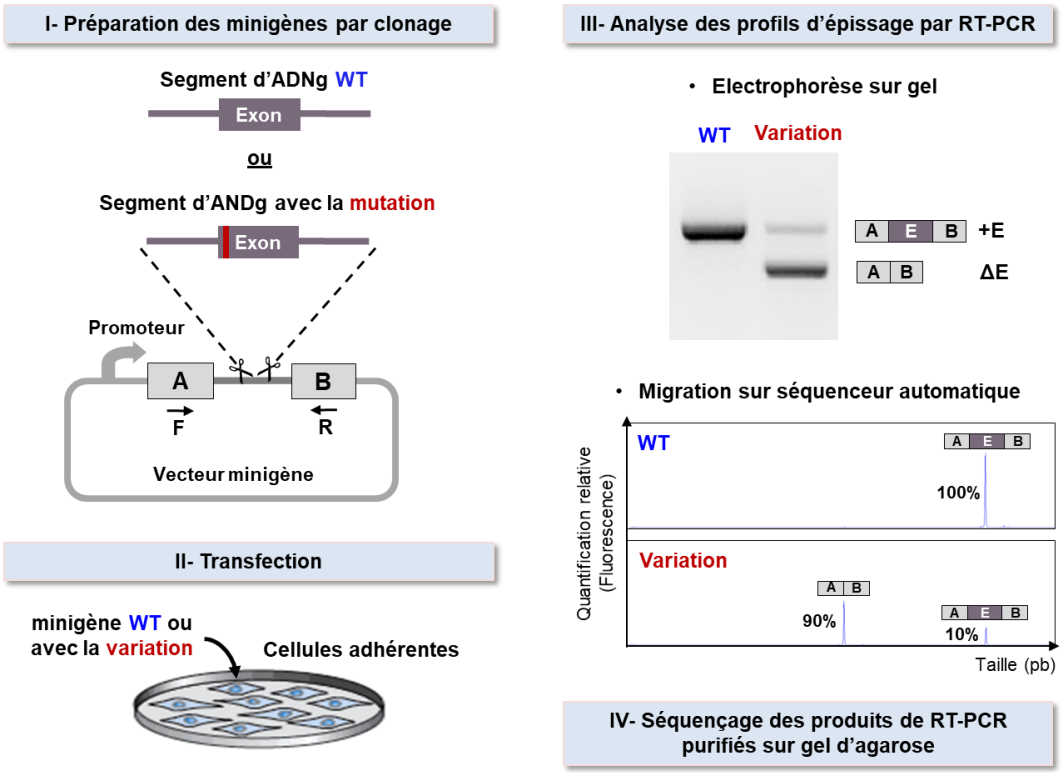
Le RNA-seq ciblé a d'ailleurs permis la caractérisation de l'épissage alternatif dans les gènes humains (Mortazavi *et al.*, 2008; Pan *et al.*, 2008; Wang *et al.*, 2009), et notamment les gènes impliqués dans le syndrome de Lynch et le syndrome seins-ovaire (Brandão *et al.*, 2019; Davy *et al.*, 2017). En effet, en comparant la proportion de *read* présents au niveau des jonctions exons-exons, représentant les différents évènements d'épissage, le RNA-seq permet potentiellement de caractériser le profil d'épissage des gènes. Cette technique présente l'avantage de pouvoir caractériser le profil d'épissage de nombreux gènes simultanément (panel de gènes ou transcriptome entier). De plus, le RNA-seq ne nécessite pas une connaissance préalable du transcriptome et permet ainsi de détecter sans *a prori* les événements d'épissage alternatifs. Cela implique que le RNA-seq peut potentiellement détecter l'existence de nouveaux transcrits correspondant à de nouvelles isoformes d'épissage, dont l'existence n'était pas soupçonnée. Aussi, le RNAseq ciblé peut permettre une analyse du transcriptome avec une résolution au niveau nucléotidique, suggérant que cette méthode est également relevante pour la détection d'éventuelles anomalies d'épissage causées par des variations génétiques insoupçonnées et l'analyse de l'expression allélique (Davy *et al.*, 2017; Wang *et al.*, 2009; pour revue : Han *et al.*, 2015).

2) Tests fonctionnels d'épissage basés sur l'utilisation de minigènes

Une autre méthode reposant sur l'utilisation, cette fois-ci, de l'ADN génomique du patient (ADNg), est aussi utilisée par les laboratoires de recherche et de diagnostic moléculaire des maladies génétiques pour analyser l'effet des variations sur l'épissage (Baralle *et al.*, 2003; pour revues : Cooper, 2005; Gaildrat *et al.*, 2010; Kishore *et al.*, 2008). Il s'agit de tests fonctionnels *ex vivo* indicateurs d'anomalies d'épissage basés sur l'utilisation de minigènes, dont le principe consiste à comparer, par RT-PCR et séquençage, le profil d'épissage des minigènes sauvage (WT, *wild-type*) et mutant, exprimés de façon transitoire dans des cellules humaines en culture (Figure 50 ; pour revues : Baralle and Baralle, 2005; Gaildrat *et al.*, 2010). Un minigène représente une version simplifiée d'un gène et correspond à un vecteur d'expression (le plus souvent un plasmide) contentant un ensemble d'exons séparés par des introns, précédé par un promoteur constitutif et suivi d'un signal de polyadénylation (pour revue : Baralle and Baralle, 2005). Bien que

classiquement les profils d'épissage soient comparés par migration des produits de RT-PCR sur gel d'agarose, il est également possible de réaliser des analyses semi-quantitatives par électrophorèse sur capillaire ou des analyses quantitatives par RT-qPCR (Figure 50 ; Hernández-Imaz *et al.*, 2015; pour revues : Baralle and Buratti, 2017; Baralle *et al.*, 2009).

Figure 50 : Principe du test fonctionnel indicateur d'anomalies d'épissage, basé sur l'utilisation de minigène. Le segment génomique d'intérêt est amplifié par PCR à partir de l'ADN génomique de patients porteurs de la variation d'intérêt puis cloné dans l'intron du minigène, entre les exons A et B, au niveau des sites de restriction. De façon alternative, les variations peuvent être introduites par mutagenèse dirigée. Après transfection dans les cellules adhérentes, le profil d'épissage des transcrits produits à partir des minigènes sauvage et muté, est déterminé par RT-PCR à l'aide des amorces sens et anti-sens respectivement localisées dans les exons A et B du minigène. Les produits de RT-PCR sont ensuite analysés par électrophorèse, comme illustré ci-dessus, et séquencés. Ici, par exemple, dans le contexte sauvage (WT, Wild Type), l'exon d'intérêt est inclus dans le transcrit mature entre l'exon A et B (+ Exon), tandis que dans le contexte muté (VAR, Variation), un saut d'exon est observé (ΔExon).

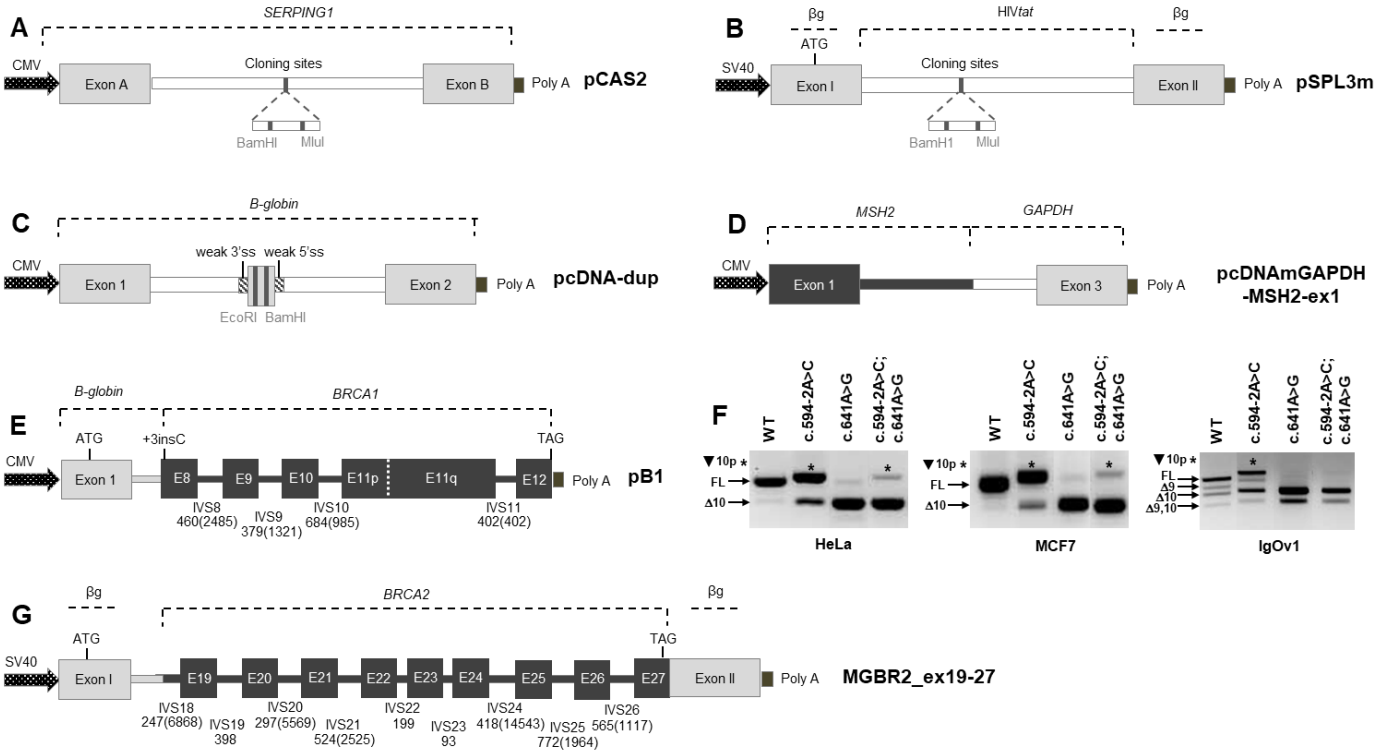


Dans le test fonctionnel *ex vivo* indicateur d'anomalies d'épissage basé sur l'utilisation de minigènes, le fragment d'intérêt contenant la variation, exonique ou intronique, est généralement amplifié à partir de l'ADN génomique du patient, puis inséré dans le vecteur d'intérêt via des sites de restrictions spécifiques (Figure 50; pour revues : Cooper, 2005; Gaildrat *et al.*, 2010). Cette méthode représente donc une alternative aux analyses réalisées à partir de l'ARN de patients, matériel biologique souvent indisponible, contrairement à l'ADN génomique du patient. De plus, lorsque l'ADNg des patients n'est pas disponible, la variation à tester peut être introduite dans le fragment d'ADN à cloner par mutagenèse dirigée avant que ce dernier ne soit inséré dans le minigène (Di Giacomo *et al.*, 2013; Ho *et al.*, 1989; van der Klift *et al.*, 2015). Les « tests minigènes » présentent également de nombreux autres avantages. Tout d'abord, contrairement à l'approche utilisant l'ARN des patients, ce système est mono-allélique et permet l'analyse allèle-spécifique des variations. Ils permettent donc d'analyser les variations d'une manière individuelle, indépendamment de la présence éventuelle d'autres variations ou de SNPs (pour revue : Fredericks *et al.*, 2015) et cela n'exclut pas pour autant la possibilité d'évaluer l'effet combiné sur l'épissage de plusieurs variations présentes sur le même allèles (en cis) soit directement à partir de l'ADNg de patients soit en les préparant par mutagenèse dirigée (de la Hoya *et al.*, 2016). De plus, contrairement à l'analyse de l'ARN des patients, cette approche utilise des contextes cellulaires et génétiques identiques entre l'analyse des profils d'épissage des constructions WT et mutante, permettant d'établir une causalité directe entre la présence d'une variation et un défaut d'épissage, le cas échéant. De même, le profil d'épissage tissu-spécifique peut être également appréhendé par transfection du minigène dans différents lignées cellulaires générées à partir de différents tissus (Figure 50 F; de la Hoya *et al.*, 2016). Enfin les transcrits produits à partir de certains minigènes comme le minigène pCAS2 (pour revue : Gaildrat *et al.*, 2010), ne sont pas dégradés par le système NMD, facilitant l'interprétation des résultats. Dans ce cas, les transcrits porteurs d'un PTC sont rendus résistants au NMD par inhibition de la traduction, via l'utilisation d'agents inhibiteurs de la synthèse protéique (puromycine ou cycloheximide, par exemple) ou bien par l'inactivation du codon d'initiation de la traduction (pour revue : Gaildrat *et al.*, 2010).

Différents types de minigènes sont décrits dans la littérature et ont été construits pour des applications bien précises. En premier lieu, les minigènes dits « universels » sont des vecteurs constitués de 2 exons séparés par un intron central qui renferme des sites de restriction au niveau desquels l'exon d'intérêt sauvage ou mutant entouré par ses séquences introniques flanquantes

(~150 nucléotides) est introduit pour former un minigène hybride à 3 exons. Il s'agit par exemple des minigènes, pCAS2 et pSPL3m, les plus utilisés notre équipe et aussi par un certain nombre de laboratoires pour la réalisation des tests fonctionnels indicateurs d'épissage, et en particulier par les laboratoires de diagnostic (Figures 51A et 51B ; Steffensen *et al.*, 2014; pour revue : Gaildrat *et al.*, 2010). En effet, ces minigènes permettent d'analyser l'effet sur l'épissage de n'importe quelle variation, située aussi bien dans l'exon que dans les régions introniques flanquantes, identifiée dans les exons cassettes (en dehors des premiers ou derniers exons). Récemment, pour pallier au manque d'outils permettant l'analyse de l'effet sur l'épissage des variations localisées dans les premiers et derniers exons, des minigènes spécifiques à l'analyse de ces exons ont d'ailleurs été développés (Figure 51D ; Naruse *et al.*, 2009 ; Inserm UMR 1245). C'est le cas notamment des minigènes pcDNAmGapdh1-MSH2-ex1 et pCAS2-MSH2ex16, valides par comparaison avec les analyses réalisées à partir des ARN de patients, mis au point pour appréhender l'effet sur l'épissage des variations situées dans les premier et dernier exon du gène *MSH2* (Naruse *et al.*, 2009 ; Inserm UMR 1245). Toutefois, ces constructions « simples » ne contenant que l'exon d'intérêt, ne permettent pas d'appréhender les profils d'épissage alternatifs complexes, comme par exemple le saut combinés de plusieurs exons. C'est pourquoi d'autres minigènes plus complexes ont été développés, minigènes avec un contexte génomique plus large qui permet de se rapprocher davantage du contexte génomique naturel et parfois appelé midigènes, lorsqu'ils contiennent la quasi-totalité du gène (Figure 51F ; Sangermano *et al.*, 2018). Ces minigènes présentent au moins 2 exons consécutifs du gènes d'intérêt (Acedo *et al.*, 2012, 2015; Baralle *et al.*, 2006; Bianchi *et al.*, 2011; Fraile-Bethencourt *et al.*, 2017; de la Hoya *et al.*, 2016b; Ramalho *et al.*, 2016; Raponi *et al.*, 2012, 2014; Sangermano *et al.*, 2018; Sharma *et al.*, 2014). C'est le cas notamment du minigène pB1, développé pour étudier l'épissage alternatif du gène *BRCA1* entre les exons 8 et 12 (Figures 51E et 51G ; Raponi *et al.*, 2012) ou des minigènes MGBR2_ex14-20 et MGBR2_ex19–27, construits pour d'étudier l'épissage alternatif du gène *BRCA2* entre les exons 14-20 et 19-27, respectivement (Figure 51G ; Acedo *et al.*, 2012, 2015; Fraile-Bethencourt *et al.*, 2017). Ces constructions peuvent contenir la totalité des séquences introniques présentes de part et d'autres des exons d'intérêt ou bien une partie seulement de ces introns alors raboutés pour permettre une taille raisonnable de l'amplicon, compatible avec le clonage (constructions pB1 ou MGBR2_ex19–27, par exemple).

<u>Figure 51 :</u> Description de vecteurs minigènes couramment utilisés dans le test fonctionnel indicateur d'anomalies d'épissage. (A) Le vecteur minigène pCAS2. (B) Le vecteur minigène pSPL3. (C) Le vecteur minigène pcDNA-dup. (D) Le vecteur minigène pcDNAm-GAPDH-MSH2-ex1. (E) Le vecteur minigène pB1. (F) Analyse du profil d'épissage alternatif du gène *BRCA1* à l'aide du minigène pB1 dans plusieurs lignées cellulaires. (G) Le vecteur minigène MGBR2_ex19-27.



De ce fait, toute variation, qu'elle soit exonique (du premier au dernier exon) ou intronique (régions introniques flanquantes et variations introniques profondes) peut théoriquement être étudiée dans des tests fonctionnels d'épissage basés sur l'utilisation de minigènes (pour revue : Gaildrat *et al.*, 2010). D'ailleurs, cette approche a permis l'identification d'un très grand nombre de mutations d'épissage localisées dans différents gènes (*CFTR*, *BRCA1/2*, MMR, *NF1*, *SMN1/2*, notamment) associés à différentes pathologies (mucoviscidose, syndrome seins-ovaires, syndrome de Lynch, Neurofibromatose de type I, SMA, respectivement) (Baralle *et al.*, 2006; Di Giacomo *et al.*, 2013; Fraile-Bethencourt *et al.*, 2017; Goina *et al.*, 2008; Hernández-Imaz *et al.*, 2015; Pagani *et al.*, 2003b, 2005; Raponi *et al.*, 2007, 2011; Singh *et al.*, 2004a, 2004b, 2007; Soukarieh *et al.*, 2016). Bien que la majorité des mutations ayant des effets sur l'épissage ait été décrites comme affectant les sites d'épissage, de plus en plus d'études, menées sur certains exons, mettent en évidence grâce à l'utilisation des minigènes, l'implication des éléments régulateurs de l'épissage

(Baralle *et al.*, 2006; Di Giacomo *et al.*, 2013; Fraile-Bethencourt *et al.*, 2017; Goina *et al.*, 2008; Hernández-Imaz *et al.*, 2015; Julien *et al.*, 2016; Ke *et al.*, 2018; Pagani *et al.*, 2003b, 2005; Raponi *et al.*, 2007, 2011; Singh *et al.*, 2004a, 2004b, 2007; Soukarieh *et al.*, 2016; Tajnik *et al.*, 2016). En effet, par exemple, l'utilisation du minigène pSPL3m a permis de montrer que 67% des substitutions ponctuelles analysées dans l'exon 10 du gène *MLH1* ont un effet sur l'épissage (Soukarieh *et al.*, 2016). De même, le minigène pTB a révélé que 76% (13/17) des variations analysées altèrent l'épissage de l'exon 5 du gène *FIX* (Tajnik *et al.*, 2016). D'ailleurs, certains minigènes ont été développés spécifiquement pour la caractérisation des régions exoniques régulatrices d'épissage. Il s'agit en particulier des minigènes pcDNA-Dup (Figure 51 ; Di Giacomo *et al.*, 2013; Soukarieh *et al.*, 2016; Tournier *et al.*, 2008) ou SXN13 (pour revue : Baralle and Baralle, 2005), utilisés dans un test dit ESE-dépendant. Ce test consiste en l'insertion d'un fragment exonique d'environ 30 pb au niveau de l'exon central alternatif du minigène dont l'inclusion dépend de la présence d'éléments régulateurs activateurs (*enhancers*) de l'épissage. En effet, l'exon central de ces minigènes étant sensible à la présence d'ESE, seuls les fragments riches en ESE seront inclus dans le produit d'épissage. Sur ce même principe, des minigènes indicateurs d'éléments inhibiteurs d'épissage (ESS) ont également été développés, tels que le minigène pZW4 (Wang *et al.*, 2004). L'utilisation de ce type de minigènes a ainsi permis la caractérisation de régions régulatrices d'épissage dans différents exons tels que l'exon 7 de *SMN1*, les exons 10 et 11 de *MLH1*, les exons 5 et 10 de *MSH2* et l'exon 7 de *BRCA2* et l'identification de plusieurs mutations de régulation d'épissage en concordance avec les résultats obtenus dans les minigènes pCAS2 et pSPL3m (Di Giacomo *et al.*, 2013; Soukarieh *et al.*, 2016; Tournier *et al.*, 2008). Avec l'essor du séquençage de nouvelle génération, les tests fonctionnels basés sur l'utilisation de minigènes ont par la suite été plus largement utilisés pour étudier à grande échelle la régulation de l'épissage et son caractère alternatif, en leur associant des stratégies de séquençage haut débit (Julien *et al.*, 2016; Ke *et al.*, 2011, 2018; Mueller *et al.*, 2015; Rosenberg *et al.*, 2015). Cette approche connue sous le nom de MPRAs (*massively parallel reporter assays*) consistant en la création de librairies contenant des millions de séquences à tester dans le système minigène a été notamment utilisée pour identifier de manière fonctionnelle des éléments de régulation de l'épissage (Ke *et al.*, 2011) et pour étudier l'effet des variations sur l'épissage via des stratégies de saturation mutationnelle dans l'exon 7 de *SMN1*, l'exon 6 de *FAS* et l'exon 5 de *WT1*, par exemple (Julien *et al.*, 2016; Ke *et al.*, 2018; Mueller *et al.*, 2015; Rosenberg *et al.*, 2015).

Malgré l'aspect artificiel d « test minigène », cette approche a été validée par de nombreux travaux, notamment ceux menés par notre laboratoire, démontrant une bonne concordance entre les données obtenues à partir des ARN des patients et celles obtenues à partir des tests fonctionnels basés sur l'utilisation de minigènes (Gaildrat *et al.*, 2012; Houdayer *et al.*, 2012; Tournier *et al.*, 2008; pour revue Baralle and Buratti, 2017). En effet, depuis 2005, notre laboratoire a développé et optimisé des tests fonctionnels d'épissage basés sur l'utilisation de différents minigènes (dont pCAS2 et pSPL3m), en principe applicables à tous les exons casettes de tout gène d'intérêt. A ce jour, plus de 1000 variations nucléotidiques ont été analysées dans les gènes BRCA et MMR (Bonnet *et al.*, 2008; Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012; Soukarieh *et al.*, 2016; Théry *et al.*, 2011; Tournier *et al.*, 2008 ; Inserm UMR 1245). Et, ces résultats ont pu être en partie validés par comparaison avec les résultats issus de l'analyse de l'ARN de patients, lorsque disponible (Gaildrat *et al.*, 2012; Soukarieh *et al.*, 2016; Tournier *et al.*, 2008), démontrant ainsi la fiabilité des minigènes comme outil de détection des mutations d'épissage. Néanmoins, quelques discordances ont été observées, notamment dans l'exon 13 de *MSH2*, ou l'effet de certaines variations induisant un saut de l'exon 13 semblerait surestimé par le minigène pCAS2, comparativement à ce qui effectivement observé par RT-PCR sur l'ARN de patients (données non publiées). D'où l'intérêt d'une validation systématique des minigènes par comparaison des résultats obtenus pour quelques variations avec ceux issus de l'analyse de l'ARN de patients. Pour autant, il n'en reste pas moins que les approches basées sur des minigènes ou sur l'ARN de patients sont des approches complémentaires qui ont permis de démontrer l'effet sur l'épissage de nombreuses variations et qui ont ainsi contribué à établir le caractère délétère d'un grand nombre de VSI. En effet, d'après les résultats obtenus sur 433 VSI identifiées par les laboratoires du GGC, il a été constaté qu'une grande fraction de ces VSI (29%) induit des modifications d'épissage, qui correspondent pour 18% d'entre elles à des effets drastiques sur l'épissage (saut total d'un exon, par exemple). Ces données ont pu être utilisées dans le cadre du diagnostic comme aide à l'interprétation et la classification des variations, avec des conséquences importantes pour une prise en charge optimale des patients et des apparentés.

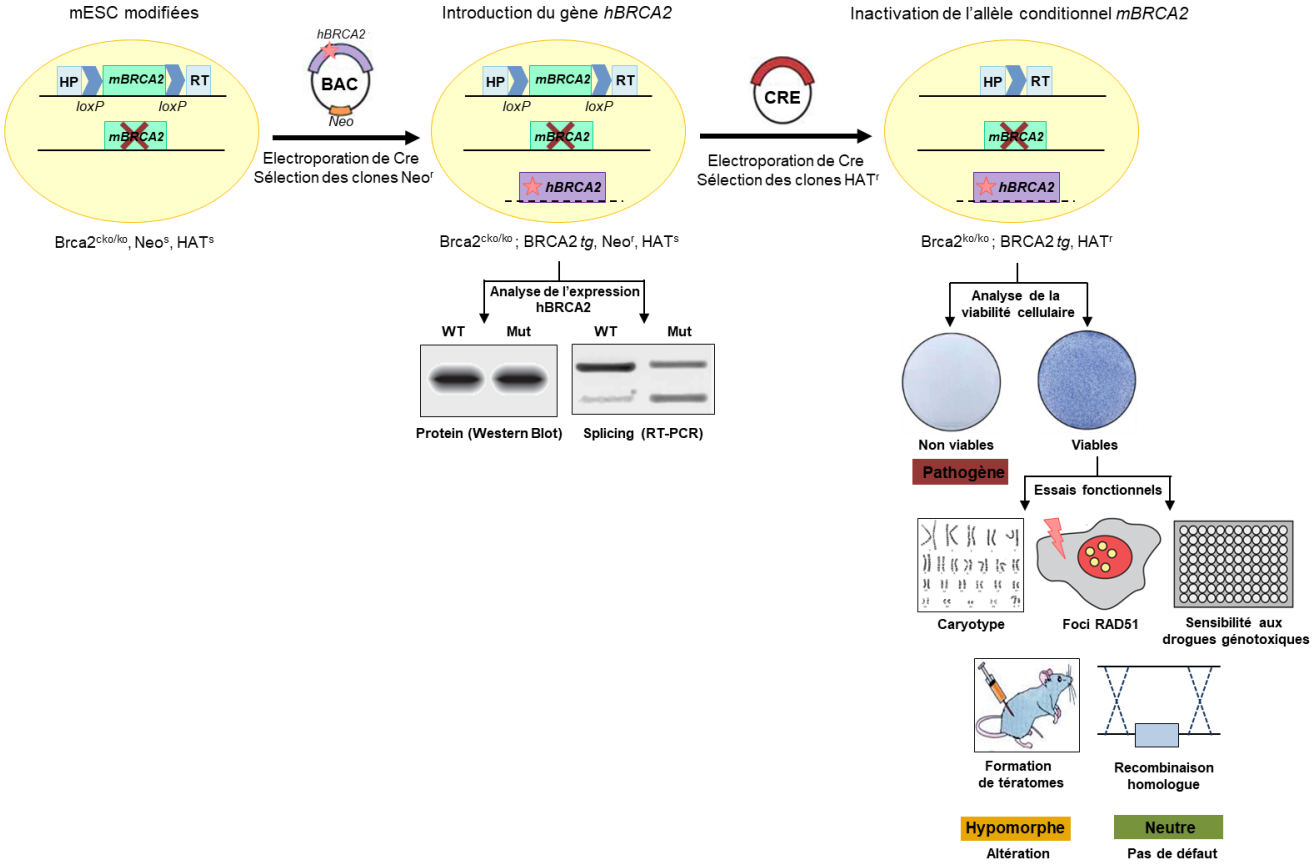
3) Essai fonctionnel basé sur l'utilisation de cellules souches embryonnaires de souris

Depuis une dizaine d'années, de nouveaux tests fonctionnels ont été spécifiquement développés pour étudier l'impact sur l'épissage des variations des gènes BRCA mais également les conséquences fonctionnelles, au niveau protéique, de telles anomalies de l'épissage, en intégrant une analyse globale de l'effet combiné de ces variations à la fois sur l'épissage et la protéine. En effet, contrairement aux tests fonctionnels indicateurs d'anomalies d'épissage basés sur l'utilisation de minigènes qui reposent sur des vecteurs d'expression portant un petite portion du gène (généralement l'exon d'intérêt entouré de ses séquences introniques flanquantes), l'essai fonctionnel basé sur l'utilisation de cellules souches embryonnaires de souris (cellules mES, *mouse embryonnic stem cells*) repose sur l'introduction du gène *BRCA1* ou *BRCA2* humain entier (séquences codantes et non codantes) sauvage ou porteur de la mutation étudiée, permettant ainsi d'appréhender l'impact de n'importe quelle variation sur l'épissage dans un contexte génomique proche du naturel (Figure 52 ; Chang *et al.*, 2009; Kuznetsov *et al.*, 2008, 2010). De plus, cet essai présente également l'avantage d'évaluer la fonctionnalité de la protéine BRCA résultante. En effet, il s'agit d'un test de complémentation visant à évaluer la capacité de transgènes humains porteurs de variations à restaurer l'activité de *BRCA1* ou *BRCA2*, tout en sachant que la viabilité des cellules mES dépend de protéines BRCA fonctionelles (Chang *et al.*, 2009; Kuznetsov *et al.*, 2008, 2010).

Il consiste à introduire, via des BAC (*bacterial artificial chromosome*), le gène *BRCA1* ou *BRCA2* humain entier sauvage ou porteur des mutations étudiées, dans des cellules mES modifiées, rendues déficientes en BRCA1 ou BRCA2 endogène de façon conditionnelle grâce au système Cre-Lox, suite à l'introduction des transgènes. La survie des cellules mES étant ainsi dépendante de l'activité BRCA1 ou BRCA2 transgénique, une variation sera considérée pathogène si elle entraine la mort cellulaire. En revanche, une variation sera considérée hypomorphe ou neutre si elle permet la viabilité cellulaire et qu'elle induit ou pas des altérations de la fonction des protéines BRCA1 ou BRCA2, respectivement. La fonctionnalité des protéines BRCA1 et BRCA2 résultantes est alors appréhendée grâce à différents essais fonctionnels basées sur l'évaluation des capacités des protéines BRCA à remplir leurs fonctions (recombinaison homologue, sensibilité aux agents génotoxiques, formation de focis RAD51, caryotypage et formation des tératomes). Ainsi, sur la base de cet essai, il est alors possible d'evaluer la pathogénicité de n'importe quelle variation des gènes BRCA1/2, aussi bien localisées dans les régions codantes que non codantes, par des

approches combinant des analyses sur ARN et protéines (Chang *et al.*, 2009; Kuznetsov *et al.*, 2008, 2010).

Figure 52 : Principe de l'essai fonctionnel basé sur l'utilisation de cellules souches embryonnaires de souris (adapté de Kuznetsov *et al.*, 2008)**.** BAC, *bacterial artificial chromosome* ; cko, *conditional knock out* ; CRE, causes recombination recombinase ; HAT, hypoxanthine aminoptérine thymidine ; h, *human* ; m, *mouse* ; mESC, *mouse embryonic stem cells* ; mut, mutant: tg, transgene ; WT, wild type.



Cette approche a déjà contribué à l'analyse de la pathogénicité d'un grand nombre de variations au sein de la plateforme d'analyse des variations BRCA pilotée par le Dr Sharan (*Center for Cancer Research & National Cancer Institute,* Frederick, Etats-Unis ; Bakker *et al.*, 2014; Biswas *et al.*, 2011, 2012; Chang *et al.*, 2009; Kuznetsov *et al.*, 2008; Li *et al.*, 2009 ; SK Sharan) et également par l'équipe du Dr Vreeswijk (LUMC, Leiden, Pays-Bas) ayant implanté cette approche depuis peu au sein de son laboratoire (Hendriks *et al.*, 2014; Mesman *et al.*, 2018;

Shimelis *et al.*, 2017). L'effet et la pathogénicité de certaines des variations analysées ont d'ailleurs pu être confirmées par des analyses à partir du matériel biologique du patient (Biswas *et al.*, 2011) ou à l'aide des données génétiques, cliniques et familiales (Mesman *et al.*, 2018; Shimelis *et al.*, 2017), validant cette approche comme outil d'aide à l'interprétation des VSI.

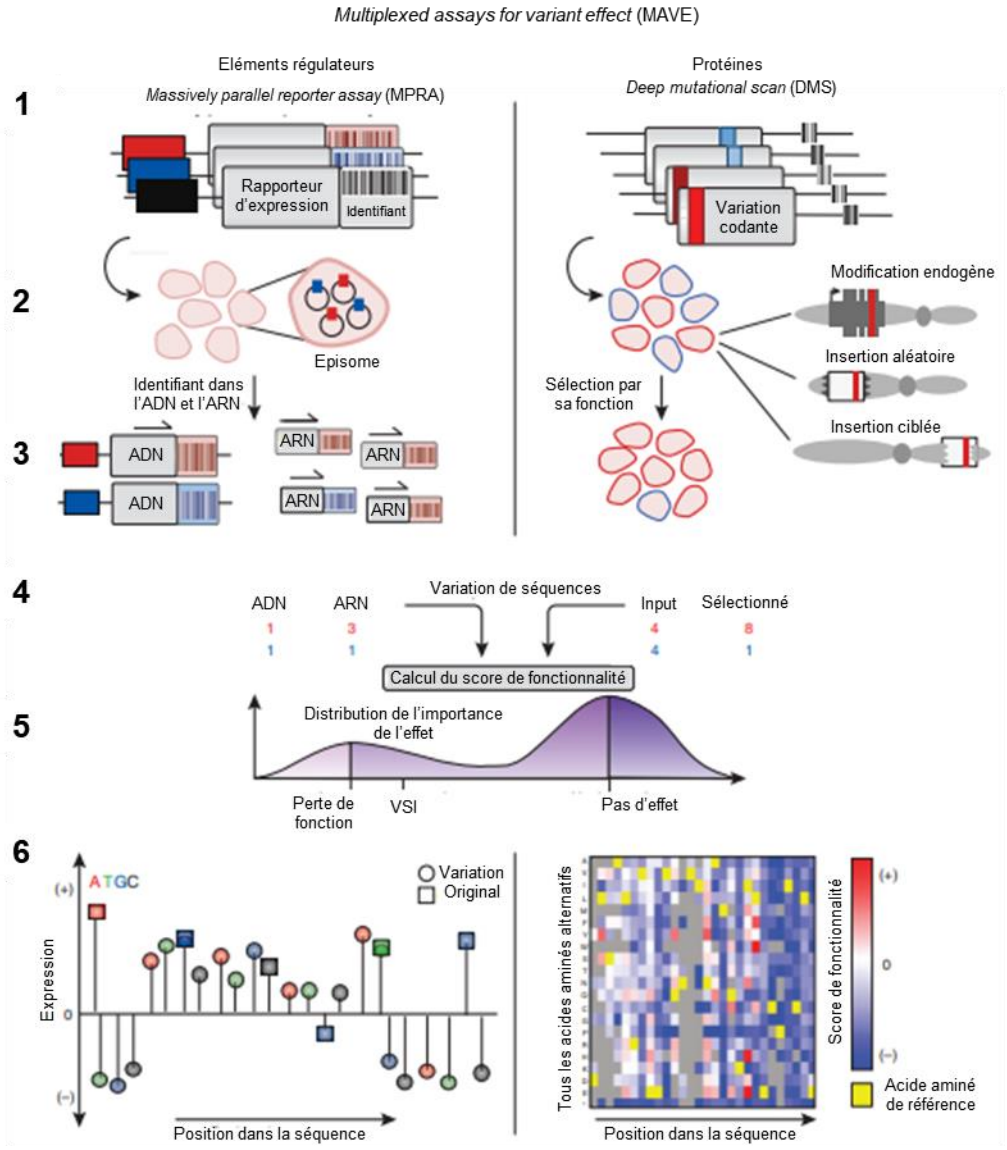
4) Essai fonctionnel basé sur la modification du génome à saturation

Les tests fonctionnels d'épissage sont traditionnellement effectués rétrospectivement, à la demande des laboratoires de diagnostic, après l'identification d'une VUS chez un patient évocateur d'une maladie génétique. La nature « une par une » de ces approches, souvent coûteuses et fastidieuses, rend celles-ci bien souvent trop lentes pour bénéficier au patient chez qui le variant a été détecté. Pour pallier à ces limitations, il a été récemment développé des approches de MAVEs (*multiplexed assays for variant effect*) consistant à collecter des données fonctionnelles pour un nombre massif de variations analysées parallèlement, en une seule et unique expérience (pour revues : Gasperini *et al.*, 2016; Starita *et al.*, 2017). Différentes stratégies de MAVEs ont été développées, mais toutes partagent un procédé commun : (i) la construction de la librairie de variations de séquence d'intérêt, (ii) l'introduction de la librairie dans le modèle d'étude (*in vitro* ou *in vivo*), (iii) l'évaluation des conséquences fonctionnelles selon un phénotype d'intérêt, (iv) le séquençage de la librairie pour quantifier la fréquence du variant avant et après sélection et (v) le calcul et la calibration des scores de fonctionnalité pour chaque variation (Figure 53 ; pour revues : Gasperini *et al.*, 2016; Starita *et al.*, 2017). Ces approches ont d'ailleurs déjà utilisées pour l'analyse séquence-fonction de nombreuses séquences régulatrices de l'expression du génome, notamment les *enhancers*, les promoteurs , les régions non-traduites (3' et 5' UTR), les sites d'épissage, les régions régulatrices de l'épissage et les régions codantes (pour revues : Gasperini *et al.*, 2016; Starita *et al.*, 2017).

Certains MAVEs ont d'ores et déjà été développés pour étudier l'impact des variations sur l'épissage, soit à partir des tests minigènes (Julien *et al.*, 2016; Ke *et al.*, 2018; Rosenberg *et al.*, 2015; Section), soit par des stratégies de modification du génome à saturation (Findlay *et al.*, 2014, 2018). Cette approche, appelée SGE (*saturation genome editing*), consiste à muter avec précision chaque nucléotide d'une séquence donnée par toutes les combinaisons de bases possibles, un

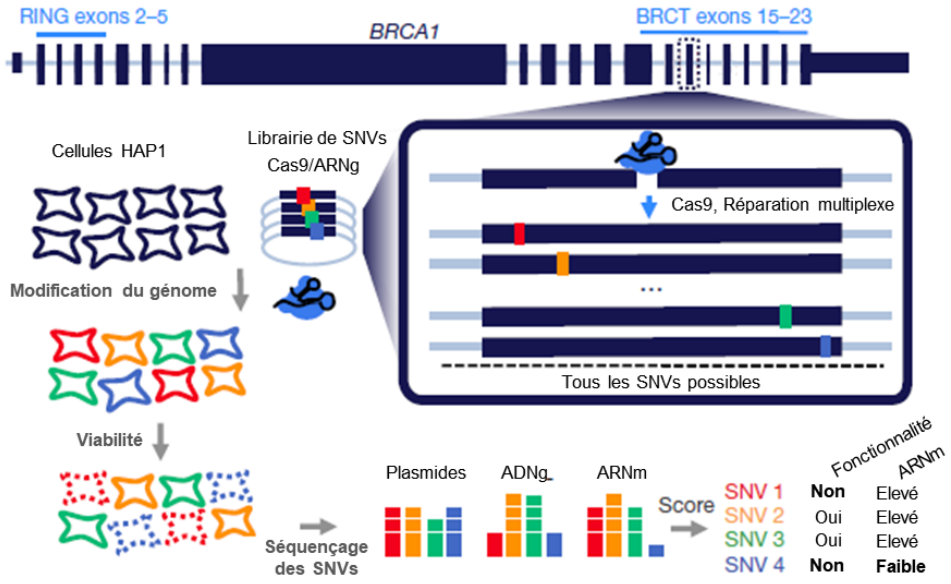
nucléotide à la fois (Findlay *et al.*, 2014, 2018). Elle présente ainsi l'avantage de pouvoir étudier l'effet combiné des variations d'intérêt sur plusieurs niveaux de la régulation de l'expression des gènes notamment l'épissage du pré-ARNm, la traduction et l'activité protéique, dans un environnement physiologique proche du contexte naturel.

Figure 53 : Principe des MAVEs (*multiplexed assays for variant effect*) utilisés pour l'analyse fonctionnelle d'un nombre massif de variations, en parallèle (d'après de Gasperini *et al.*, 2016)**.**



Récemment, cette approche a été appliquée à l'étude de la fonctionnalité de nombreuses variations (~3,893) localisées dans le gène *BRCA1* (Findlay *et al.*, 2018). Cette protéine contient les domaines RING et BRCT, essentiels pour son activité, codés par les exons 2 à 5 et 15 à 23, respectivement, et pour lesquels il a été démontré l'existence de variations délétères à l'origine de la perte de fonction du gène. Chacune des positions nucléotidiques de ces 13 exons a été mutée, via la technique de CRISPR/Cas9, dans les cellules haploïdes HAP1 (1 exon pour 20 millions de cellules HAP1) dont la viabilité est dépendante de l'intégrité de la voie de réparation par recombinaison homologue (HDR) dans laquelle sont impliqués des gènes essentiels, parmi lesquels les gènes de prédisposition au syndrome seins-ovaires (*BRCA1/2*, *RAD51B/C/D*, *PALB2*, *BARD1)* (Blomen *et al.*, 2015; Findlay *et al.*, 2018). Après 11 jours de culture, l'ADNg a été séquencé afin d'appréhender la fréquence à laquelle chaque variation était présente au sein de la population cellulaire, par rapport à celle de la librairie initialement transfectée. Sur la base de ces résultats, chaque variation a pu être classée comme (i) fonctionnelle, si la fréquence de la mutation dans la population cellulaire est similaire à celle observée dans la librairie initialement transfectée, indiquant que la fonction HDR est active dans les cellules porteuses de cette variation, (ii) non-fonctionnelle, si la fréquence est moins élevée que la moyenne, indiquant que le variant conduit à la mort cellulaire des cellules porteuses de cette variation ou (iii) intermédiaire (Figure 54 ; Findlay *et al.*, 2018). Ces données fonctionnelles ont pu être validées, au moins en partie, à l'aide de données cliniques disponibles pour certaines variations dans la base de données ClinVar, montrant une concordance dans plus de 96% des cas (Findlay *et al.*, 2018). De plus, l'impact de chacune de ces variations sur l'épissage a pu être appréhendé par séquençage de l'ARN, révélant que des variations exoniques associées à une déplétion en ARNm comparativement à l'ADNg affectait l'épissage de cet exon (Findlay *et al.*, 2018).

Figure 54 : Principe du SGE (*saturation genome editing*) utilisé pour introduire toutes les SNVs possibles à travers 13 exons de *BRCA1* codant pour les domaines RING (exons 2-5) et BRCT (exons 15-23) (d'après de Findlay *et al.*, 2018)**.**



CHAPITRE VII : PREDICTIONS BIOINFORMATIQUES DE L'EFFET DES VARIATIONS SUR L'EPISSAGE

En théorie, toute variation nucléotidique, qu'elle soit intronique ou exonique, est potentiellement susceptible d'affecter l'épissage de l'ARNm, soit en modifiant les signaux consensus d'épissage, soit en altérant des éléments de régulation. En réalité, la majorité des variations identifiées au niveau des gènes MMR ou BRCA ne fait pas l'objet d'une analyse fonctionnelle permettant d'évaluer leur impact sur l'épissage. Etant donné le grand nombre de VSI dans ces gènes, il est donc essentiel aujourd'hui de sélectionner, de manière rationnelle, les variations à analyser en priorité par des approches expérimentales. Cette stratification des variations s'appuie principalement sur l'utilisation d'outils bio-informatiques de prédictions des altérations de sites d'épissage alors que les éléments de régulation sont encore peu caractérisés et manquent d'outils de prédiction suffisamment performants (Houdayer *et al.*, 2012; Soukarieh *et al.*, 2016; pour revue : Spurdle *et al.*, 2008).

1) Approches bioinformatiques dédiées aux sites d'épissage

Les séquences consensus des sites 5' et 3' d'épissage étant aujourd'hui relativement bien définies, différents outils *in silico* de prédiction, notamment SpliceSiteFinder-like (SSF-L) (Shapiro and Senapathy, 1987), MaxEntScan (MES) (Yeo and Burge, 2004), Neural Network Splice (NNSplice) (Reese *et al.*, 1997) et Human Splicing finder (HSF) (Desmet *et al.*, 2009), pour ne citer que les outils les plus largement utilisés, ont été développés dans le but de prédire la position et la force des sites d'épissage potentiels, dans une séquence donnée (pour revue : Spurdle *et al.*, 2008). Bien que la majorité des outils de prédiction dédiés aux sites d'épissage soient basés sur les informations acquises sur les séquences consensus des sites 5' et 3' d'épissage, ces derniers gérèrent pourtant des scores de prédictions différents (Houdayer *et al.*, 2012). En effet, la comparaison de ces logiciels entre eux a révélé deux différences fondamentales. D'une part, chacun de ces outils résulte de l'application d'algorithmes différents, et d'autre part, chacun prend une définition des sites donneur et accepteur consensus d'épissage qui leur est propre :

- les scores SSF [0 – 100] sont générées par des matrices dites PWM (*position weight matrix*) elles-mêmes établies à partir d'un jeu de données constitué de jonctions exons/introns constitutives pour les sites donneurs et accepteurs. Ces scores reflètent la fréquence attribuée, par rapport à ce jeu de données, à chacun des nucléotides présents à chaque position des sites d'épissage (Shapiro and Senapathy, 1987). Il existe une version modifiée de SSF, SSF-like, qui permet le calcul des scores des sites donneurs d'épissage à la fois pour les motifs canoniques GT et GC.
- les scores HSF [0 – 100] sont également obtenus suite à l'utilisation des matrices PWM développées par Shapiro et Senapathy (Desmet *et al.*, 2009).
- les scores MES [0 – 16, selon l'interface Alamut Visual], ceux-ci sont basés sur la méthode d'entropie maximale qui, en utilisant un large jeu de données de sites d'épissage humains, étudie la distribution des séquences en fonction d'un ensemble de contraintes, définies par la position des nucléotides et par l'interdépendance des nucléotides adjacents et non-adjacents au sein de ces sites (Yeo and Burge, 2004).
- les scores NNSsplice [0 – 1] sont calculés quant à eux suivant une méthode de probabilité basée sur des réseaux neuronaux artificiels, l'un pour les 3'ss et l'autre pour les 5'ss, qui permet d'identifier les séquences correspondant aux sites d'épissage potentiels (Reese *et al.*, 1997).

Néanmoins, pour chacune de méthode, plus la valeur du score du site d'épissage prédit est grande au sein de chaque gamme, plus la probabilité qu'il s'agisse effectivement d'un vrai site d'épissage est élevée (Houdayer *et al.*, 2012). Les outils de prédiction dédiés aux sites d'épissage permettent d'appréhender la force de chaque site d'épissage mais également de comparer ces scores dans les contextes sauvage et mutant (pour revue : Hartmann *et al.*, 2008). De plus, ces outils ne se restreignent pas seulement à la prédiction d'altérations de la force des sites d'épissage consensus ou alternatifs naturels situés aux jonctions exon-intron de référence mais aussi permettent également de prédire la création de site d'épissage de *novo* et la position et force de sites d'épissage cryptiques exoniques ou introniques (Desmet *et al.*, 2009; Houdayer *et al.*, 2012). Sur la base du changement de la force des sites d'épissage induits par des variations, ces outils permettent aujourd'hui de prédire si ces variations sont susceptibles ou pas d'induire des anomalies d'épissage (pour revue : Hartmann *et al.*, 2008).

Plusieurs études ont démontré une très bonne concordance entre ce type de prédictions et les altérations de l'épissage effectivement observées expérimentalement pour de nombreuses variations nucléotidiques, notamment pour les gènes MMR, BRCA, *RB1* et *CFTR* (Houdayer *et al.*, 2012; Sharma *et al.*, 2014; Sharp *et al.*, 2004; Théry *et al.*, 2011; Tournier *et al.*, 2008). En conséquence, les outils de prédiction des modifications des sites d'épissage sont désormais utilisés en routine, dans les laboratoires de diagnostic moléculaire, pour sélectionner les variations qui feront l'objet en priorité d'une analyse fonctionnelle expérimentale (pour revues : Hartmann *et al.*, 2008; Spurdle *et al.*, 2008). Cette utilisation dans le cadre du diagnostic a été facilitée par le fait que ces méthodes, en libre accès, peuvent facilement être interrogées indépendamment sur des interfaces web. Mais leur implémentation dans les laboratoires de diagnostic a été facilitée par le développement, par la société *Interactive Biosoftware* (http://www.interactive-biosoftware.com), de l'interface graphique Alamut Visual qui permet, variant par variant, d'interroger simultanément l'ensemble de ces logiciels et puis plus récemment, d'Alamut Batch, logiciel d'annotation (notamment pour les prédictions d'épissage) à haut débit destiné aux analyses NGS.

Toutefois, très peu d'études ont comparé les prédictions obtenues par ces différents outils de prédictions et évalué leur fiabilité c'est-à-dire la capacité des approches à discriminer les variations qui provoquent un effet sur l'épissage de celles qui n'en provoquent pas (Houdayer *et al.*, 2012; Leman *et al.*, 2018; Sharma *et al.*, 2014; Tournier *et al.*, 2008). Plus spécifiquement, l'étude rétrospective menée en 2012 par le groupe de travail « GGC-épissage » sur 272 variations situées dans les gènes *BRCA1* et *BRCA2*, a permis de démontrer une meilleure capacité prédictive pour MES suivi par SSFL, par comparaison des scores prédictifs obtenus pour chacune des variations avec les résultats expérimentaux obtenus par différents laboratoires du réseau d'oncogénétique français (Houdayer *et al.*, 2012). D'ailleurs, d'autres études prospectives menées ont confirmé l'excellente capacité de prédiction de MES (Leman *et al.*, 2018; Moles-Fernández *et al.*, 2018; Sharma *et al.*, 2014; pour revue : Hartmann *et al.*, 2008).

Ces travaux ont également permis d'établir et de recommander des seuils pour l'utilisation de chacun de ces logiciels, notamment de 15% pour les scores MES et de 5% pour les scores SSFL, à utiliser de manière séquentielle (Houdayer *et al.*, 2012). Ces seuils, qui correspondent à la valeur de la différence entre les scores des sites d'épissage dans les contextes sauvage et mutant à partir de laquelle un effet sur l'épissage est observé, représentent le meilleur compromis entre spécificité

et sensibilité (Houdayer *et al.*, 2012). En effet, jusqu'à cette étude et en l'absence de recommandations spécifiques, le seuil à partir duquel tout changement de scores obtenus par les approches *in silico* dédiés aux sites d'épissage était considéré comme significatif, c'est-à-dire prédictif d'un défaut d'épissage, était souvent choisi de manière aléatoire (Bonnet *et al.*, 2008; Houdayer *et al.*, 2012). L'ensemble de ces données ont finalement permis à Houdayer et ses collaborateurs de proposer, pour la première fois, des recommandations pour l'analyse bio-informatique des variations qui affectent les sites consensus d'épissage, telles que les variations qui entraînent une diminution de 15% du scores MES, suivie par une diminution de 5% du score SSF-L sont celles qui devraient être analysées en priorité à l'aide de tests fonctionnels d'épissage (Houdayer *et al.*, 2012).

Plus récemment, il a été montré qu'une nouvelle stratégie combinant les données issues de ces 2 outils serait encore plus performante pour la prédiction des anomalies d'épissage car basée sur une régression logistique représentant la meilleure combinaison possibles des scores MES et SSFL (Leman *et al.*, 2018). Ainsi en 2018, à nouveau grâce à l'effort collaboratif du « GGC-épissage » joint cette fois-ci au consortium ENIGMA, un nouvel outil de prédiction axés sur les sites d'épissage a pu être développé et validé sur 395 variations localisées dans les gènes *BRCA1/2* (305 variations) et dans 9 autres gènes (90 variations réparties sur les gènes *CFTR, CTRC, HFE, HJV, LRP5, PKD1, RHD, SLC40A1 et TFR2*) (Leman *et al.*, 2018). Etant donné les très bonnes performances de cet outil (~99% des défauts d'épissage ont été correctement prédits), le protocole SPiCE (*Splicing Prediction in Consensus Elements*) est aujourd'hui proposé comme une nouvelle recommandation à suivre pour la prédiction des variations affectant potentiellement les régions consensus des sites d'épissage (donneur: positions -3 à +8 et accepteur : positions -12 à +2) (Leman *et al.*, 2018). Et cet outil, disponible en libre accès, peut facilement être implémenté dans les laboratoires de diagnostics. En revanche, il est important de noter qu'à l'heure actuelle, il n'existe encore aucune recommandation concernant les variations situées en dehors des régions consensus des sites d'épissage.

2) Approches bioinformatiques dédiées aux points de branchement

D'autres outils bio-informatiques, ceux-ci basés sur la séquence consensus yUnAy, qui caractérise la majorité des sites de branchement (BS, *branch site*), ont été conçus pour prédire la

position et/ou l'altération potentielle de ces sites, notamment SROOGLE (Schwartz *et al.*, 2009) ou Human Splicing Finder (Desmet *et al.*, 2009). Cependant, aucun de ces outils n'a fait l'objet de véritables études visant à évaluer leur fiabilité, par comparaison des données *in silico* qu'ils génèrent avec celles obtenues expérimentalement. Ainsi, aujourd'hui, les BPs humains restent très difficiles à prédire en se basant uniquement sur la séquence consensus, étant donné la dégénérescence élevée de la séquence autour du BP (Gao *et al.*, 2008). De plus, seules quelques centaines de BPs ont été identifiés chez l'homme et seulement une vingtaine de mutations pathogènes à l'origine de défauts d'épissage par altération des BP ont été identifiées (Bitton *et al.*, 2014; Gao *et al.*, 2008; Královicová *et al.*, 2006; Taggart *et al.*, 2012; pour revue : Lewandowska, 2013). A cela, s'ajoute le fait que certains introns possèdent des multiples points de branchement, parfois même éloignés du site 3' d'épissage (BPs distants) ou bien sans l'adénosine canonique au niveau du nucléotide de branchement (Bitton *et al.*, 2014; Gao *et al.*, 2008; Taggart *et al.*, 2012). Malgré leur importance dans le processus d'épissage et leur implication dans des maladies génétiques, la localisation et les caractéristiques des points de branchement restent à l'heure actuelles très largement mal connues (Mercer *et al.*, 2015; pour revues : Padgett, 2012; Singh and Cooper, 2012).

Plus récemment et grâce au développement des techniques de séquençage a haut-débit, des études à large échelle ont été menées sur les BPs (Chiang *et al.*, 2017; Corvelo *et al.*, 2010; Mercer *et al.*, 2015; Signal *et al.*, 2018; Taggart *et al.*, 2012, 2017; Zhang *et al.*, 2017). Ces études, pour la plupart, visaient à caractériser l'ensemble des points de branchement du génome humain, par des approches de *mapping*, c'est-à-dire par une identification des points de branchement à l'échelle du génome entier. Ces données ont ensuite été compilées et intégrées par des algorithmes d'apprentissage automatique (*maching learning*) de manière à générer des outils d'annotations de points de branchement et des outils de prédictions de l'impact de SNV sur les points de branchement tels que BPP (*branch point predictions*) ou SVM-BPfinder (*support vector maching learning for branch points*) et Branch Pointer (Corvelo *et al.*, 2010; Mercer *et al.*, 2015; Signal *et al.*, 2018; Zhang *et al.*, 2017). Bien que non exhaustives, ces analyses à large échelle ont non seulement permis d'identifier et caractériser des points de branchement mais ont également permis de confirmer l'importance biologique d'un certain nombre de ces sites lorsqu'ils sont altérés par des variations nucléotidiques. Ils nécessitent maintenant d'être évalués de façon élargie avant de

pouvoir être implémenté dans les laboratoires de diagnostic pour la stratification des variations à analyser en priorité dans les tests fonctionnels.

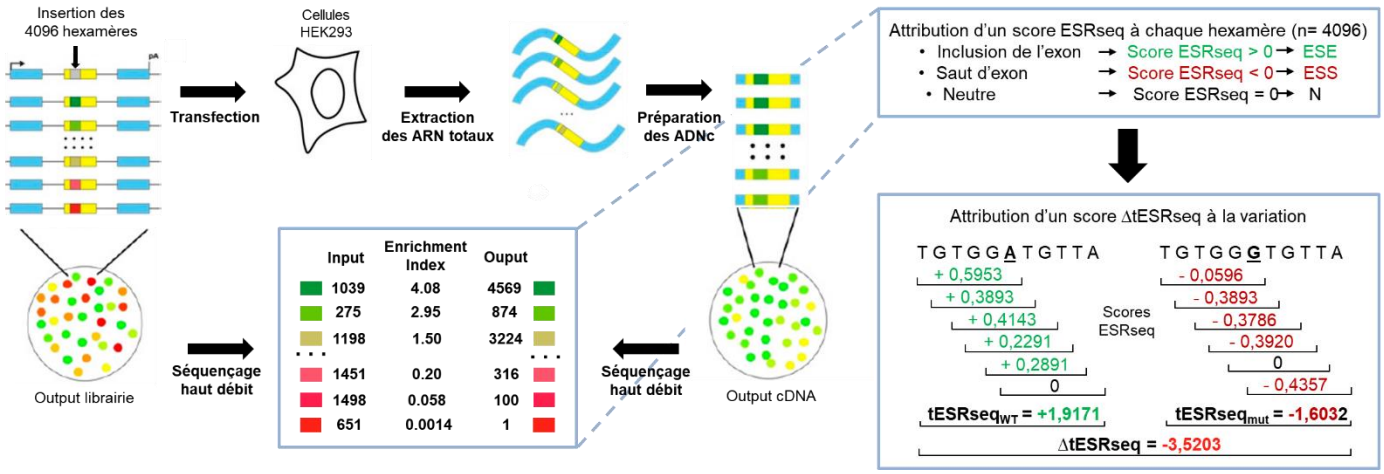
3) Approches bioinformatiques dédiées aux éléments cis régulateurs d'épissage

De nombreuses approches, telles que RESCUE-ESE, ESE-finder, HSF, EX-SKIP parmi tant d'autres, ont également été développées afin de prédire l'existence, dans une séquence donnée, d'éléments de type ESE, ESS, ISE ou ISS participant à la régulation de l'épissage (Cartegni *et al.*, 2003; Desmet *et al.*, 2009; Fairbrother *et al.*, 2002; Raponi *et al.*, 2011). Ces différentes approches, intégrées dans différents outils bioinformatiques, sont dérivées de méthodes statistiques et/ou expérimentales et plus précisément sur (i) l'analyse de l'enrichissement des motifs proches des sites d'épissage (Castle *et al.*, 2008; Fairbrother *et al.*, 2002; Zhang and Chasin, 2004), (ii) les données de conservation des séquences entre les espèces (Goren *et al.*, 2006) et (iii) sur les séquences régulatrices d'épissage identifiées expérimentalement (Cartegni *et al.*, 2003; Piva *et al.*, 2009, 2012; Smith *et al.*, 2006). Il apparait donc possible, en théorie, de prédire au moyen de ces outils, l'impact d'une variation sur les éléments régulateurs de l'épissage.

Cependant, l'interprétation de ce type de prédictions reste encore très difficile (Spurdle *et al.*, 2008). En effet, du fait de la multiplicité des matrices dégénérées intégrées dans ces outils, de très nombreux motifs sont identifiés dans une séquence donnée, motifs souvent chevauchants et non exhaustifs. Plusieurs travaux ont montré que les prédictions générées par ce type d'outils sont peu fiables et ne peuvent être utilisés pour la stratification des variations pour les tests fonctionnels d'épissage (Arnold *et al.*, 2009; Lastella *et al.*, 2006; Raponi *et al.*, 2007; Soukarieh *et al.*, 2016; Tournier *et al.*, 2008). En effet, les éléments de régulation d'épissage sont encore très mal définis et pas complètement caractérisés, compliquant ainsi leur prédiction (Spurdle *et al.*, 2008). De plus, leur mode d'action combinatoire et chevauchant, dépend de plusieurs paramètres tels que leur distance par rapport aux sites d'épissage, leur position et interdépendance par rapport à d'autres séquences régulatrice de l'épissage ou la structure secondaire de l'ARN (Fairbrother *et al.*, 2004). En conséquence, la génération d'outils *in silico* efficaces pour prédire les altérations d'éléments régulateurs d'épissage reste donc un enjeu majeur.

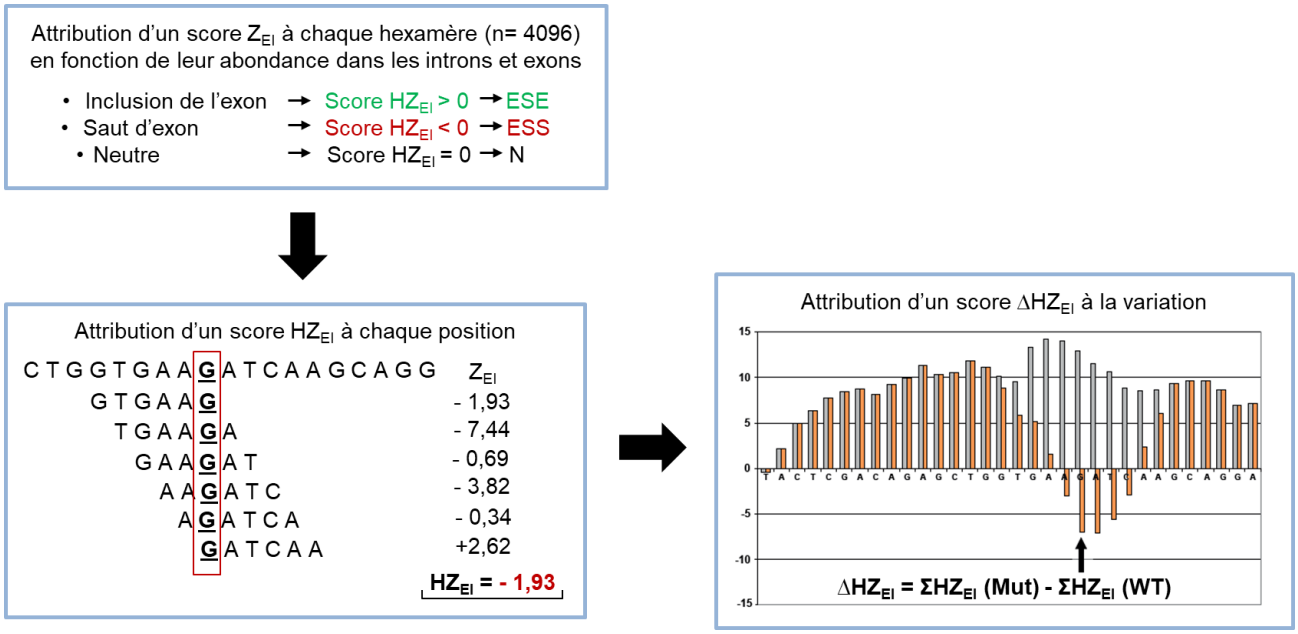
Récemment, quatre nouvelles méthodes, dérivées de méthodes statistiques et/ou expérimentales, ont été décrites comme outils très prometteurs pour la prédiction des variations affectant potentiellement des éléments régulateurs de l'épissage. Tout d'abord, la première approche repose sur des travaux ayant pour objectif l'identification fonctionnelle et globale d'éléments exoniques régulateurs de l'épissage, de type ESE et ESS (Ke *et al.*, 2011). Ils ont contribué à évaluer l'effet sur l'épissage de tous les motifs possibles de 6 nucléotides (4096 hexamères au total) dans un contexte exonique. Plus précisément, en associant des tests fonctionnels basés sur l'utilisation de minigènes, et des stratégies de séquençage haut débit, ces travaux ont permis de quantifier la capacité de chaque hexamère à induire l'inclusion ou l'exclusion d'exons cassettes (Figure 55). Sur la base de ces résultats, un score ESRseq (*exonic splicing regulator sequence*) compris entre -1 et +1 a ensuite été attribué à chacun des 4096 hexamères : un score ESRseq positif, négatif ou nul correspond, respectivement, à un élément potentiel de type ESE, ESS ou neutre. Il est alors théoriquement possible d'appréhender l'impact d'une variation exonique sur la régulation de l'épissage en calculant la valeur du changement des scores ESRseq totaux ($\Delta tESR_{seq}$, *total ESRseq score change*) induit par la variation, par rapport au contexte sauvage, en prenant en considération les 6 hexamères chevauchant la position de la variation d'intérêt (Di Giacomo *et al.*, 2013).

Figure 55 : Prédictions des altérations d'éléments régulateurs de l'épissage selon la méthode QUEPASA (adapté de Ke *et al.*, 2011**).**



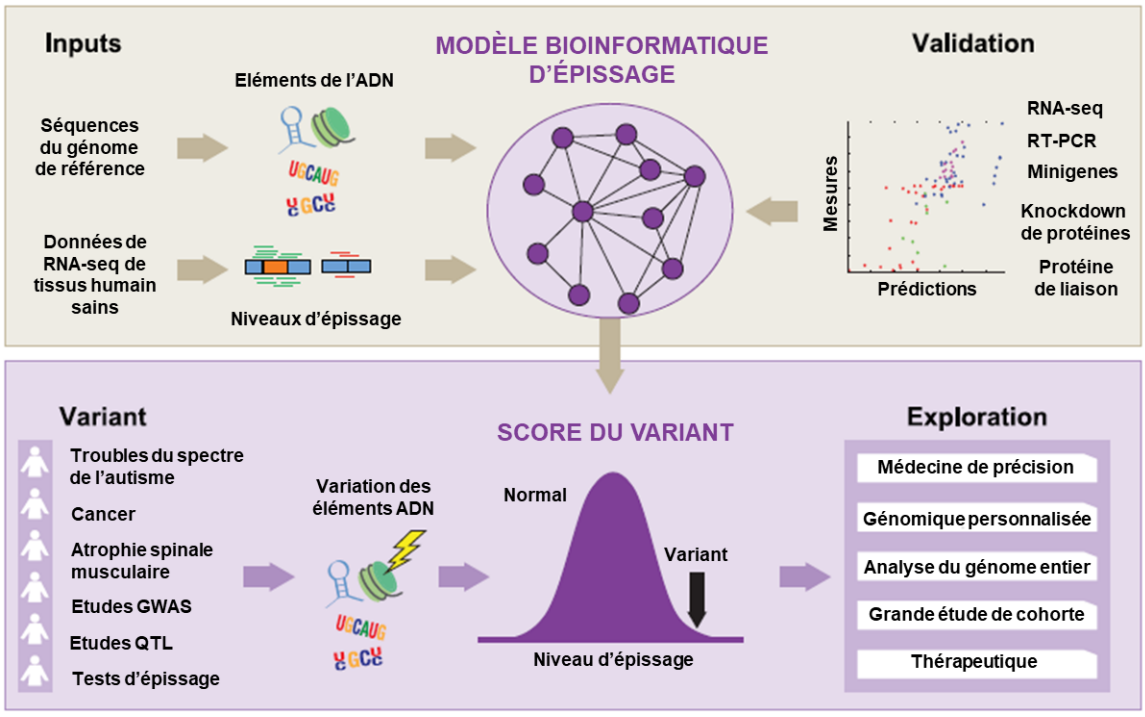
La deuxième approche HEXplorer dérive d'une étude de type RESCUE (Fairbrother *et al.*, 2002), combinée à la notion selon laquelle la reconnaissance du site d'épissage serait plus fortement dépendante de la présence d'ESR dans le cas de sites d'épissage dits faibles comparativement aux sites d'épissage forts (Erkelenz *et al.*, 2014). Ainsi, les hexamères associés aux ESE sont généralement supposés être surreprésentés (i) dans les séquences exoniques situées en amont des 5' ss faibles, comparativement aux séquences localisées en amont des 5' ss forts, de même que (ii) dans les séquences exoniques versus les séquences introniques. Plus simplement, Erkelenz et ses collaborateurs ont évalué la distribution relative des hexamères dans les exons et les introns en fonction de leur distance au site d'épissage (Erkelenz *et al.*, 2014). A partir de cette analyse statistique, un score Z_{EI} a été attribué à chacune des combinaisons d'hexamères (Figure 56). Ainsi, comme pour le calcul $\Delta tESR_{seq}$, l'effet potentiel d'une variation exonique sur les éléments régulateurs de l'épissage peut être déterminé par le calcul du score ΔHZ_{EI}, correspondant à la différence du total des scores HZ_{EI} des 6 hexamères chevauchants la position de la variation d'intérêt entre l'allèle sauvage et l'allèle muté (Erkelenz *et al.*, 2014 ; https://www2.hhu.de/rna/html/hexplorer _score. php).

<u>Figure 56 :</u> Prédictions des altérations d'éléments régulateurs de l'épissage selon la méthode HEXplorer.



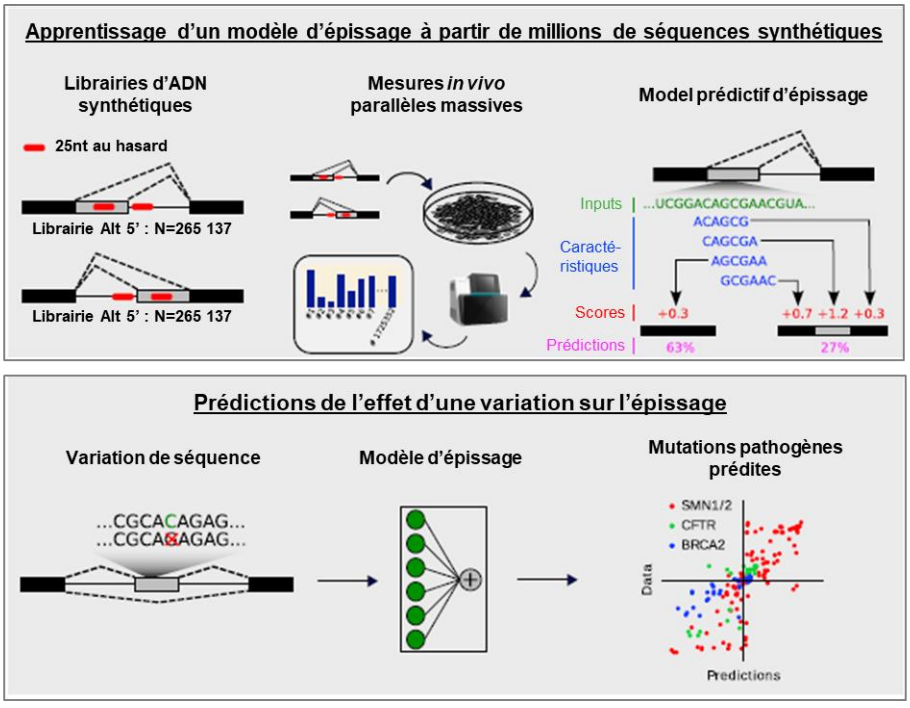
La troisième approche, SPANR (*splicing-based analysis of variants*), repose sur une méthode d'apprentissage automatique (*maching learning*) corrélant 1393 caractéristiques extraites de la séquence d'ADN de l'exon d'intérêt aux données d'épissage obtenues par RNAseq à partir de 16 tissus différents humains sains (Figure 57 ; Xiong *et al.*, 2015). A partir de l'ensemble de ces données, un modèle computationnel d'épissage a été généré afin de prédire le pourcentage de transcrit contenant l'exon (PSI ou Ψ, *percent spliced in*) pour chacun des 16 tissus humains, à la fois dans les contextes sauvage (génome de référence) et muté. L'outil rapporte ensuite le changement maximum du PSI induit par la mutation dans les 16 tissus par rapport au contexte sauvage (ΔΨ-SPANR). Cette approche permet ainsi de prédire l'effet de toutes substitutions exoniques et des substitutions introniques localisées jusqu'à 300 nucléotides par rapport aux sites d'épissage, situées au niveau des éléments de régulation mais aussi au niveau des sites d'épissage.

Figure 57 : Prédictions des altérations d'éléments régulateurs de l'épissage selon le modèle computationnel d'épissage SPANR (d'après Xiong *et al.*, 2015)**.**



La quatrième et dernière approche, HAL (*hexamer additive linear*) est un modèle de prédiction reposant sur une combinaison d'approches de biologie synthétique et d'apprentissage automatique en utilisant les profils d'épissage générés à partir de 2 librairies constituées dans leur ensemble de près de 2 millions de minigènes synthétiques épissés alternativement (Figure 58 ; Rosenberg *et al.*, 2015). Plus précisément, cette approche a consisté à mesurer par séquençage parallèle massif via la technique de RNA-seq les ratios des différentes isoformes générées par épissage alternatif à partir des millions de minigènes synthétiques, notamment le pourcentage de transcrit contenant l'exon (PSI). Les données générées ont ensuite été utilisées pour entraîner un model prédictif d'épissage alternatif qui convertie les caractéristiques des 6-mer en un score prédictif (ΔΨ-HAL), comparatif des contextes sauvage et muté.

Figure 58 : Prédictions des altérations d'éléments régulateurs de l'épissage selon le modèle prédictif d'épissage alternatif (HAL) tiré de million de séquences synthétiques (d'après Rosenberg *et al.*, 2015)**.**



A ce jour, une seule étude a procédé à l'évaluation et la comparaison du pouvoir prédictif de trois de ces quatre nouveaux outils de prédictions axés sur les ESR ($\Delta tESR_{seq}$, ΔHZ_{EI} et $\Delta\Psi$-SPANR) (Soukarieh *et al.*, 2016). Cette étude a démontré une très bonne concordance entre ces nouvelles méthodes de prédictions basées sur les ESR, au moins en ce qui concerne les scores $\Delta tESR_{seq}$ et ΔHZ_{EI}, et les altérations de l'épissage observées expérimentalement pour les variations nucléotidiques issues de cinq jeux de données (*MLH1* exon 10, *BRCA2* exon 7, *BRCA1* exon 6, *CFTR* exon 12 et *NF1* exon 37) comparativement aux outils précédents tels que HSF-SR, EX-SKIP et ESEfinder (Soukarieh *et al.*, 2016). Au-delà de prédire quelles variations exoniques affectent l'épissage, les scores $\Delta tESR_{seq}$ et ΔHZ_{EI} sont également capables de prédire la direction (corrélation entre le niveau d'inclusion de l'exon et les valeurs des scores prédictifs) et la sévérité (séparation des variations en 3 groupes : les variations qui augmentent le saut d'exon *versus* les variations sans effet sur l'épissage *versus* les variations qui augmentent l'inclusion de l'exon) du défaut d'épissage. En revanche, les scores $\Delta\Psi$-SPANR n'ont pas montré un pouvoir prédictif convaincant (Soukarieh *et al.*, 2016).

Ces nouvelles approches bio-informatiques semblent donc être prometteuses pour la détection des mutations qui altèrent potentiellement la régulation de l'épissage et pourraient représenter une nouvelle stratégie de stratification des variations à analyser expérimentalement. Il est donc essentiel de poursuivre l'évaluation de ces méthodes de façon plus générale afin de mieux déterminer leur fiabilité, leur applicabilité et surtout leurs limitations, avant de les proposer comme des outils de stratification des variations génétiques pour des tests fonctionnels d'épissages au titre du diagnostic, non seulement dans des gènes impliqués dans le syndrome de Lynch ou dans le syndrome seins-ovaires (modèles d'étude de ce projet de thèse), mais aussi dans d'autres pathologies d'origine génétique.

CHAPITRE VIII : INTERPRETATION BIOLOGIQUE DES VARIATIONS ASSOCIEES A DES DEFAUTS D'EPISSAGE

1) Classification des variations selon leur effet sur l'épissage

Ces dernières années, de nombreux travaux, et en particulier ceux menés par notre unité, ont permis de mettre en évidence, à l'aide de tests fonctionnels, l'impact sur l'épissage d'un grand nombre de variations dans les gènes MMR et BRCA (Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012; Houdayer *et al.*, 2012; Soukarieh *et al.*, 2016; Théry *et al.*, 2011; Tournier *et al.*, 2008). Depuis 2012, ces variations étudiées pour leur impact potentiel sur l'épissage, au niveau de l'ARN, sont classées selon leur effet en 3 classes, de 1S à 3S (Houdayer *et al.*, 2012) :

- La classe 1S regroupe les variations sans effet sur l'épissage analysées par minigène ou par RT-PCR. Dans le cas d'une analyse par RT-PCR, cela nécessite (i) d'utiliser des inhibiteurs du NMD et (ii) de confirmer la présence des 2 allèles pour les variations exoniques et d'utiliser un SNP exonique informatif pour les variations introniques.
- La classe 2S fait référence aux variations qui ont un effet sur l'épissage mais un effet partiel ou un renforcement d'épissage alternatif. Dans ce cas, l'allèle muté produit deux transcrits, un transcrit pleine longueur et un autre transcrit aberrant ou alternatif.
- La classe 3 englobe les variations qui ont un effet total sur l'épissage, c'est-à-dire que l'allèle mutant ne produit pas du tout de transcrit pleine longueur.

Mais cette classification au niveau de l'ARN ne conditionne pas nécessairement la classification clinique d'une variation qui dépend plutôt de l'impact de l'épissage anormal observé sur la fonctionnalité de la protéine. En effet, les variations sans effet sur l'épissage (classe 1S) peuvent être (i) non pathogènes (classe 1) s'il s'agit de variations synonymes ou introniques *a priori* sans effet sur la fonctionnalité de la protéine, (ii) pathogènes (classe 5) en ce qui concerne les variations non-sens ou *frame-shift* et (iii) de signification inconnue (classe 3) pour les variations faux-sens, en raison de leur impact potentiel sur la protéine.

Dans la majorité des cas, les variations des gènes BRCA et MMR pour lesquelles un effet total sur l'épissage (classe 3S) a été montré, conduisent à des défauts d'épissage hors phase ayant pour conséquence un décalage du cadre de lecture dans la séquence codante et l'introduction d'un

PTC. Le transcrit correspondant alors porteur d'un PTC devient cible du NMD et peut être potentiellement dégradé par celui-ci ou traduit en une protéine tronquée non fonctionnelle. Ces variations sont donc responsables de la perte de fonction du gène et peuvent être par conséquent classées pathogènes (classe 5) (Walker *et al.*, 2013). Il en est de même pour les variations qui sont responsables d'un défaut d'épissage en phase, conservant le cadre de lecture, mais conduisant à une protéine avec une délétion interne qui affecte un domaine fonctionnel important (Walker *et al.*, 2013). C'est le cas de certaines variations localisées dans l'exon 3 de *BRCA2*, à l'origine d'un saut total en phase de l'exon 3. Cet exon est essentiel à l'activité de la protéine BRCA2, puisqu'il contient (i) le domaine de transactivation avec lequel interagit EMSY, un régulateur négatif de l'activité transcriptionnelle de BRCA2 et (ii) le domaine d'interaction avec la protéine PALB2 qui fait le lien entre BRCA1 et BRCA2 (Martinez *et al.*, 2015). Cette interaction est d'ailleurs essentielle au recrutement de BRCA2 au niveau des DSBs et à la recombinaison homologue (Hartford *et al.*, 2016; pour revue : Prakash *et al.*, 2015). En effet, il a été montré que des variations faux-sens qui perturbent cette interaction (W31R et W31C) altèrent également la fonction de BRCA2 et notamment la réparation des DSBs par recombinaison homologue, rendant les cellules sensibles aux dommages de l'ADN (Biswas *et al.*, 2012). De plus, des variations à l'origine du saut total de l'exon 3, tout comme la délétion génomique de l'exon 3, sont associées à une augmentation du risque de développer un cancer du sein et/ou de l'ovaire et ont été classées pathogènes sur la base de données de co-ségrégation et de données fonctionnelles (Caputo *et al.*, 2018; Muller *et al.*, 2011b).

2) Problématique de l'interprétation des variations associées des défauts d'épissage en phase

Certaines variations provoquent des altérations de l'épissage entrainant des modifications en phase de la protéine, en dehors d'une région codant pour un domaine fonctionnel protéique. Les conséquences de ces anomalies d'épissage sur la fonctionnalité des protéines résultantes demeurent inconnues. Il est alors impossible de statuer sur leur caractère neutre ou délétère. De telles variations restent donc classées comme VSI (classe 3) et ne peuvent être utilisées au titre du diagnostic. De même, certaines variations sont localisées au niveau de régions exoniques épissées alternativement en phase. Ces transcrits alternatifs, dont les niveaux ne sont pas nécessairement augmentés par les variations, contiennent des délétions internes en phase conduisant

potentiellement à la production de protéines fonctionnelles. Les variations localisées dans ces exons restent alors considérées comme des VSI (Tableau 8 ; Walker *et al.*, 2013). C'est le cas tout particulièrement des variations localisées les exons 9 et 10 de *BRCA*1 et des exons 4-7 et 12 de *BRCA2*.

Les exons 9 et 10 de *BRCA1* codent pour une région de la protéine BRCA1 sans fonction connue. Il a été démontré que l'isoforme BRCA1 Δ9-10 est produite de manière physiologique (Colombo *et al.*, 2014). Il a également été montré que la variation c.594-2A>C en cis avec c.641A>G (c.[594-2A>C ;641A>G]), initialement classée pathogène, n'était en fait pas associée à une augmentation du risque de développer un cancer, bien que ce variant soit responsable du saut hors phase de l'exon 10, sans altérer de façon significative le niveau de Δ9,10 (de la Hoya *et al.*, 2016b; Rosenthal *et al.*, 2015). En effet, il a été suggéré que l'isoforme Δ9-10 pourrait être fonctionnelle et permettre un mécanisme de sauvetage ou d'atténuation du phénotype, sans que cela n'ait jamais été démontré de manière fonctionnelle (de la Hoya *et al.*, 2016b). De même, des travaux menés sur l'exon 12 de *BRCA2*, naturellement alternativement épissé, ont montré à l'aide de tests fonctionnels basés sur l'utilisation de cellules souches que la délétion en phase de cet exon, entraînant la délétion interne de 32 acides aminés (p.Glu2282_2313del), n'altère pas la fonction de la protéine BRCA2, ce qui pourrait s'expliquer par le fait que cette région ne correspond à aucun domaine fonctionnel connu (Li *et al.*, 2009a). Ces travaux ont également montré que le variant *BRCA2* c.6853A>G (p.Ile2285Val) était responsable d'une augmentation du saut de cet exon. Malgré cet effet sur l'épissage, ce variant a pu être classé non pathogène sur la base, d'une part, de l'absence de co-ségrégation de ce variant avec la maladie dans les familles de porteurs, et, d'autre part, du fait de la co-occurrence en trans d'une mutation pathogène chez certains individus porteurs de c.6853A>G (Li *et al.*, 2009a). L'ensemble de ces données suggère la redondance fonctionnelle de l'exon 12 de *BRCA2* (Li *et al.*, 2009a). D'autres travaux menés cette fois-ci sur l'exon 7 de *BRCA2* ont également montré à l'aide de tests fonctionnels basés sur l'utilisation de cellules souches, que les variations c.631+2T>G et c.581G>A (p.W194X) supposées pathogènes, étaient à l'origine du saut de l'exon 7 majoritairement, mais conduisaient également à la production d'un transcrit alternatif Δ4-7 très minoritaire. Ce transcrit, restaurant la phase, code pour une protéine délétée dans sa partie interne de 105 acides aminés mais fonctionnelle, au moins au niveau de son activité de réparation de l'ADN (Biswas *et al.*, 2011). Ces données ont ensuite été confirmées (i) *in vivo*, à partir d'un model murin *BRCA2 knock-in* dans lequel les exons 4 à 7 ont été délétés,

montrant que ces exons n'étaient pas nécessaires à la viabilité et ne modifiaient pas la survie sans tumeur (Thirthagiri *et al.*, 2016), et (ii) du fait de l'identification de la variation c.631+2T>G, en co-occurrence en trans avec une mutation pathogène, chez des individus atteints d'une anémie de Fanconi, ce qui est *a priori* non compatible avec la viabilité de ces individus si la perte de fonction est totale et les deux allèles sont tous les deux nuls (la déficience en BRCA2 étant létale au stade embryonnaire) (Biswas *et al.*, 2011; Sharan *et al.*, 1997).

Il est possible que certaines des variations systématiquement considérées pathogènes, en particulier les variations type non-sens ou de type insertions/délétions à l'origine d'un décalage du cadre de lecture ou encore les variations introniques localisées au niveau des positions invariables AG/GT (-2,-1/+1,+2) des sites d'épissage (pour revue : Baralle *et al.*, 2009), pourraient conduire à la production d'une protéine fonctionnelle, au moins en partie, par le biais d'une modification en phase de l'épissage, i.e. le saut de l'exon 12, le saut des exons 4-7 ou le saut des exons 9-10 (Tableau 8). Même si démontré fonctionnellement, il serait important de confirmer l'existence possible d'un mécanisme de sauvetage par les isoformes BRCA2 Δ4-7 et Δ12 et BRCA1 Δ9-10 à l'aide d'analyses complémentaires, basées notamment sur la collecte des données génétiques, cliniques, tumorales et familiales des patients porteurs afin de réévaluer le caractère pathogène de ce type de variations. Si cette hypothèse s'avère correcte, cela représenterait un nouveau paradigme dans le cadre du syndrome seins-ovaires. Ces résultats seraient alors susceptibles de remettre en cause la classification de certaines mutations pathogènes et pourraient conduire au déclassement de certaines variations de *BRCA1* ou *BRCA2* actuellement considérées pathogènes. Mais au vue des connaissances actuelles ces variations doivent être considérées comme VSI, jusqu'à preuve du contraire (Tableau 8).

Jusqu'à lors, seule une variation initialement supposée pathogène dans *BRCA2* a fait l'objet d'une reclassification en variation non causale. Il s'agit de la variation polymorphique non-sens *BRCA2* c.9976A>T (p.Lys3326*), à l'extrémité C-terminale de la protéine, entraînant seulement la délétion des 93 acides aminés C-terminaux (Mazoyer *et al.*, 1996; Meeks *et al.*, 2016). Cependant, dans d'autres pathologies, quelques exemples de mutations supposées pathogènes (mutations non-sens et sur les positions les plus conservées des sites consensus d'épissage) ont été décrites comme associées à une atténuation du phénotype, par des modifications en phase de

Tableau 8 : Prudence dans l'interprétation des variations *BRCA1* et *BRCA2* prédites comme pouvant augmenter le niveau de transcrits alternatifs qui contiennent des délétions internes en phase conduisant potentiellement à la production de protéines BRCA fonctionnelles. (adapté de de la Hoya *et al.*, 2016).

Gène	Evénement d'épissage alternatif	Variation d'intérêt	Rationnel
BRCA1	Δ8p	c.442-1 (IVS7-1) c.442-2 (IVS7-2)	Le site accepteur de l'exon 8 de *BRCA1* correspond à un site accepteur d'épissage en tandem (NAGNAG) sujet à un épissage alternatif (Colombo *et al.*, 2014). Les variations en position c.442-1,-2 sont prédites pour détruire le site accepteur de référence (5') mais pas le site accepteur alternatif (3'), résultant en la production de transcrits Δ8p.
	Δ9,10	c.548-1,-2 (IVS8-1,-2), c.593 to non-G c.593+1,+2 (IVS9+1,+2) c.594-1,-2 (IVS9-1,-2) c.670 to non-G c.670+1,+2 (IVS10+1,+2)	Les patients porteurs d'une variation à ces positions sont prédites pour générer un taux normal (ou augmenté) de transcrits Δ9,10, événement d'épissage alternatif majeur en phase (Colombo *et al.*, 2014). La variation *BRCA1* c.594-2A>C (en co-occurrence en cis avec la variation c.641A>G) a été rapportée pour être associée à des caractéristiques cliniques incohérentes avec un risque élevé de cancer normalement associé à une variation pathogène dans *BRCA1* (Rosenthal *et al.*, 2015). L'haplotype c.[594-2A>C; 641A>G] a été montré, à partir d'analyses sur ARN de patients, pour produire des niveaux importants de transcrits Δ10 (~70% des transcrits détectés) et a été classée comme non délétère (Classe 1) par le consortium ENIGMA à l'aide d'une analyse multifactorielle incluant des données génétiques (ségrégation, analyse cas-contrôles) et cliniques (de la Hoya *et al.*, 2016b).
	Δ11, Δ11q	c.4096 to non-G c.4096+1 (IVS11+1) c.4097+2 (IVS11+2) c.4097+2 (IVS11+3)	Les données collectées par le consortium ENIGMA démontrent que les variations *BRCA1* c.4096+1G>A et c.4096+3A>G augmentent la production de Δ11q, transcrits alternatifs physiologiques en phase et également de Δ11 (Bonatti *et al.*, 2006; Byrjalsen *et al.*, 2017; Radice, non publiées). Alors que la variation c.4096+1G>A n'est associée à des traits cliniques caractéristiques des variations *BRCA1* pathogènes (Spurdle, données non publiées), la variation c.4096+3A>G a été classée probablement neutre (classe 2), sur la base de données de co-ségrégation et la présence de cette variation à l'état homozygote chez un porteur sain (Byrjalsen *et al.*, 2017). Des mutations non-sens homozygotes dans *BRCA1* exon 11 (c.1115G>A et c.1151T>G) ne sont ni associées à une augmentation du risque de développer un cancer du sein ni à une anémie de Fanconi via l'existence du transcrit alternatif physiologique Δ11q (Seo *et al.*, 2018).
	Δ13p	c.4186-1 (IVS12-1) c.4186-2 (IVS12-2)	Le site accepteur de l'exon 13 de *BRCA1* correspond à un site accepteur d'épissage en tandem (NAGNAG) sujet à un épissage alternatif (Colombo *et al.*, 2014). Les variations en position c.4186-1,-2 sont prédites pour détruire le site accepteur de référence (5') mais pas le site accepteur alternatif (3'), résultant en la production de transcrits Δ13p.
	Δ14p	c.4358-1 (IVS13-1) c.4358-2 (IVS13-2)	Le site accepteur de l'exon 14 de *BRCA1* correspond à un site accepteur d'épissage en tandem (NAGNAG) sujet à un épissage alternatif (Colombo *et al.*, 2014). Les variations en position c.4358-1,-2 sont prédites pour détruire le site accepteur de référence (5') mais pas le site accepteur alternatif (3'), résultant en la production de transcrits Δ143p.
BRCA2	Δ4-7	c.317-1,-2 (IVS3-1,-2) c.425+1,+2 (IVS4+1,+2) c.426-1,-2 (IVS4-1,-2) c.475+1,+2 (IVS5+1,+2) c.476-1,-2 (IVS5-1,-2) c.517+1,+2 (IVS6+1,+2) c.517-1,-2 (IVS6-1,-2) c.631+1,+2 (IVS7+1,+2)	Les patients porteurs d'une variation en position -1,-2 et +1,+2 des exons 4, 5, 6 et 7 sont susceptibles de générer un taux anormal de transcrits Δ4-7, qui résulte d'un événement d'épissage alternatif physiologique en phase (Biswas *et al.*, 2011; Fackenthal *et al.*, 2016). Le Δ4-7 génère une isoforme fonctionnelle (viabilité et activité de recombinaison homologue dans le test de complémentation basé sur l'utilisation des cellules mES) et qui n'est pas associée à une diminution de la survie chez la souris (Thirthagiri *et al.*, 2016). Les variations *BRCA2* c.631+2T>G et c.581G>A (p.W194X) classées au préalable comme pathogène par perte de fonction sont associées à la production d'une protéine BRCA2 fonctionnelle dans son activité de recombinaison homologue via la production d'un transcrit alternatif minoritaire délété des exons 4 à 7 (Biswas *et al.*, 2011).
	Δ12	c.6842-1,-2 (IVS11-1,-2) c.6937 to non-G c.6937+1,+2 (IVS12+1,+2)	Les variations en position c.6842-1,-2 et c.6937,-1,-2 sont prédites pour augmenter l'expression de *BRCA2* Δ12, un transcrit alternatif physiologique en phase (Fackenthal *et al.*, 2016). L'exon 12 de BRCA2 est redondant d'un point de vue fonctionnel (Li *et al.*, 2009a).
Note – D'autres variations systématiquement considérées délétères (non-sens ou *frameshift*) survenant dans les exons 4-7 et 12 de *BRCA2* et dans les exons 9 et 10 de *BRCA1* pourraient ne pas être associés à un risque élevé de cancer du sein et/ou de l'ovaire en raison de la production d'un transcrit alternatif physiologique en phase qui permet de contourner des codons stop prématurés, encodant ainsi une protéine, certes tronquée, mais fonctionnelle. L'analyse de nombreuses données cliniques (de patients et de contrôles) ne fournit pas, à l'heure actuelle, un soutien suffisamment fort à cette hypothèse (Spurdle, de la Hoya, données non publiées). Des recherches supplémentaires sont en cours pour étudier plus en profondeur l'importance fonctionnelle/clinique de ces variations.			

l'épissage, parmi lesquelles la dystrophie musculaire de Duchenne (*DMD*), la dystrophie musculaire congénitale (*LAMA2*), l'épidermolyse bulleuse simple généralisée sévère (*COL17A1*), et l'amaurose congénitale de Leber (*CEP290*) (Di Blasi *et al.*, 2001; Disset *et al.*, 2006; Flanigan *et al.*, 2011; Hinzpeter *et al.*, 2010; Kowalewski *et al.*, 2016; Littink *et al.*, 2010; Miro *et al.*, 2015; Tuffery-Giraud *et al.*, 2005).

3) Problématique de l'interprétation des variations induisant des défauts d'épissage partiels

D'autres mutations d'épissage identifiées à partir d'analyses fonctionnelles ont un effet partiel (classe 2S). Comme pour les mutations à l'origine d'anomalies d'épissage en phase en dehors de domaines protéiques connus pour leur fonction, ces mutations sont considérées comme VSI (classe 3). En effet, tandis qu'il est admis que la perte d'expression totale d'un allèle (classe 3S) peut être considérée comme délétère (Houdayer *et al.*, 2012), il demeure une question en suspens quant au seuil à partir duquel un effet partiel peut être considéré comme délétère (seuil de pathogénicité). C'est le cas notamment des variations affectant l'épissage de l'exon 3 du gène *BRCA2*. En effet, bien que cet exon soit indispensable à l'activité de la protéine BRCA2 et que des mutations associées au saut total de cet exon sont considérées comme délétères (Biswas *et al.*, 2012; Caputo *et al.*, 2018; Hartford *et al.*, 2016), il existe pourtant, de manière physiologique, une proportion faible de transcrits alternatifs de *BRCA2* sans l'exon 3 détectée dans les tissus normaux, incluant les tissus mammaires et prostatiques (Fackenthal *et al.*, 2016; Zou *et al.*, 1999). De plus, certaines variations à l'origine d'un saut très partiel de l'exon 3, en particulier la variation c.68-7T>A augmentant très partiellement le saut de l'exon 3, ne sont pas associées à une augmentation du risque de développer un cancer du sein (OR 1.03) et sont considérées comme neutres (Colombo *et al.*, 2018). Il devrait donc exister un seuil de pathogénicité à partir duquel une mutation entrainant une production trop importante de transcrit Δ3 serait délétère. La détermination de ce seuil permettrait de contribuer à l'interprétation des variations associées à des effets d'épissage partiels dans cet exon. Cette question s'applique d'ailleurs à l'ensemble des variations induisant des défauts d'épissage partiels non seulement dans les gènes BRCA mais aussi dans les gènes MMR ou dans d'autres gènes impliqués dans des maladies héréditaires.

L’évaluation de la pathogénicité des variations génétiques représente un des principaux défis actuels de la génétique médicale (Richards *et al.*, 2015). En particulier, l’interprétation biologique de l’ensemble des variations impactant l’épissage reste aujourd’hui très délicate, rendant parfois très difficile la classification clinique définitive de ces variations, potentiellement utilisables pour le conseil génétique, en particulier parce qu’elle dépend des connaissances, détenues à un moment donné. L’ensemble des travaux mentionnés soulignent l’importance de multiplier les approches complémentaires pour parvenir à la classification clinique d’une variation, qui dépend d’un faisceau d’arguments, basés sur des données génétiques, cliniques, familiales et fonctionnelles. Ces arguments ne peuvent être obtenus que grâce aux efforts déployés à plusieurs niveaux (national et international) par les laboratoires de diagnostic et de recherche qui souvent interagissent pour collecter l’ensemble de ces informations. Depuis quelques années, une grande partie des données ainsi répertoriées est partagée via des bases de données accessibles par internet. Ces interfaces numériques sont devenues cruciales pour le progrès des connaissances en génomique médicale et ne cessent de s’améliorer (Brookes and Robinson, 2015; Cline *et al.*, 2018; Johnston and Biesecker, 2013; Landrum and Kattman, 2018).

Partie II : Objectifs des travaux de thèse

Depuis l'implémentation du séquençage à haut-débit en diagnostic, l'enjeu majeur de la génétique médicale n'est plus de détecter les variations présentes dans le génome d'un patient mais plutôt d'identifier, parmi les milliers de variations détectées, celles potentiellement à l'origine de la maladie. La problématique d'interprétation des variations est devenue rapidement, avec l'augmentation exponentielle des tests génétiques réalisés à visée diagnostic, la question prioritaire des laboratoires de génétique médicale assurant le diagnostic moléculaire des maladies Mendéliennes. En effet, une des limites majeures du diagnostic moléculaire des maladies Mendéliennes est la détection d'un nombre croissant de variations nucléotidiques de signification biologique inconnue (VSI) qui, à la lumière des connaissances actuelles, restent inutilisables en diagnostic. La problématique d'interprétation des variations est particulièrement importante en oncogénétique et notamment pour le syndrome de Lynch (gènes MMR) et le syndrome seins-ovaires (gènes BRCA), deux des prédispositions héréditaires au cancer les plus fréquentes, compte-tenu : (i) du nombre important d'analyses moléculaires réalisées, (ii) de l'impact médical du résultat de l'analyse génétique et (iii) du pourcentage exceptionnellement élevé de VSI identifiées dans les gènes MMR et BRCA (~30-50% des variations répertoriées dans les bases de données), faisant de ces gènes de véritables paradigmes dans ce domaine.

Ces dernières décennies, la relecture de la complexité de l'épissage de l'ARNm a permis de mettre en évidence l'implication d'anomalies de l'épissage dans des maladies génétiques, y compris le syndrome de Lynch et le syndrome seins-ovaires. Par conséquent, certaines VSI identifiées dans les gènes MMR et BRCA pourraient être délétères du fait de l'altération de l'épissage. Dans ce contexte, les travaux de recherche de l'unité Inserm U1245 ont démontré, grâce à des tests fonctionnels basés sur l'utilisation de minigènes, qu'une proportion importante de VSI (~30% de la totalité des variations MMR et BRCA analysées au sein de notre groupe), est à l'origine de défauts d'épissage. Si la plupart des mutations d'épissage identifiées correspondent à des altérations des sites d'épissage, une minorité touche des éléments régulateurs d'épissage potentiels. Récemment, les travaux menés au sein de notre unité ont permis de mettre en évidence à l'aide d'analyses mutationnelles extensives ciblées sur des exons modèles, parmi lesquels les exons 7 et 18 de *BRCA2* (Di Giacomo *et al.,* 2013; Gaildrat *et al.,* 2012; Gaildrat *et al.,* données non publiées), l'exon 10 de *MLH1* (Soukarieh *et al.,* 2016) et l'exon 3 de *SMN1* (données non publiées), qu'il existerait, parmi les très nombreuses variations identifiées dans les gènes MMR et

BRCA, une fraction importante de variations responsables d'une altération des éléments régulateurs de l'épissage.

En réalité, la majorité des variations identifiées au niveau des gènes MMR ou BRCA ne fait pas l'objet d'une analyse fonctionnelle systématique permettant d'évaluer leur impact sur l'épissage. Etant donné le grand nombre de VSI dans ces gènes, il serait important de sélectionner, de manière rationnelle, les variations à analyser en priorité par des tests fonctionnels d'épissage, souvent chronophages et coûteux. Aujourd'hui, cette stratification des VSI s'appuie principalement sur l'utilisation d'outils bioinformatiques de prédiction des altérations de sites d'épissage alors que les éléments de régulation restent encore peu caractérisés et manquent d'outils de prédiction suffisamment fiables et performants. Dans ce contexte, nos récents travaux suggèrent que des nouvelles approches (Introduction – Chapitre VII.3) pourraient permettre de prédire l'impact de certaines VSI sur des éléments régulateurs d'épissage (ESR), et de stratifier les analyses fonctionnelles (Soukarieh *et al.*, 2016; données non publiées). Cependant, cette étude pilote n'a été réalisée que sur un jeu de données relativement petit (n=154 variations) correspondant à des données expérimentales obtenues sur cinq exons ciblés (Soukarieh *et al.*, 2016). De plus, des données additionnelles obtenues dans d'autres exons et/ou sur d'autres gènes ont révélé quelques cas discordants (Inserm UMR 1245, données non publiées). Il est donc essentiel de poursuivre l'évaluation de ces nouvelles approches de façon plus générale et à plus large échelle afin de mieux déterminer leurs performances ainsi que leur applicabilité et leurs limitations avant de les proposer, au titre du diagnostic, comme outils de stratification des variations génétiques identifiées dans des gènes impliqués dans le syndrome de Lynch ou dans le syndrome seins-ovaires mais aussi dans toute autre pathologie d'origine génétique (Résultats – Chapitre I).

De manière plus spécifique, les données fonctionnelles d'épissage obtenues au sein de notre laboratoire dans le cadre de son activité de recherche et de diagnostic sur les gènes MMR et BRCA nous ont permis d'identifier des exons modèles correspondant à des cas extrêmes de concordance-discordance entre les prédictions bioinformatiques de mutations de régulation d'épissage et les données expérimentales issues des tests fonctionnels d' épissage. Plus précisément, il s'agit (i) d'exons pour lesquels les prédictions d'altérations d'ESR potentielles ont été particulièrement concordantes avec les données expérimentales obtenues sur le petit échantillon de variations nucléotidiques testées et (ii) d'exons pour lesquels, au contraire, les nouveaux outils

bioinformatiques axés sur les ESRs ne semblaient pas performants sur le petit nombre de variations analysées. Des études mutationnelles extensives réalisées sur ces exons modèles devraient permettre d'identifier des caractéristiques spécifiques qui confèrent, à ces exons, une sensibilité ou une résistance particulière aux mutations ESR, permettant (i) d'expliquer la concordance/discordance des résultats expérimentaux avec les prédictions *in silico* et les limitations et (ii) de contribuer à l'amélioration d'outils bioinformatiques dédiés aux ESR (Résultats – Chapitre II).

L'ensemble des données fonctionnelles d'épissage obtenues au sein de notre laboratoire, contribuent, dans le cadre de collaborations nationales (GGC) et internationales (consortiums ENIGMA et InSiGHT), à l'interprétation et la classification des variations analysées dans les gènes MMR et BRCA. Cependant, si la plupart des mutations d'épissage identifiées au sein de notre unité provoque des défauts d'épissage hors phase, permettant de classer certaines de VSI comme pathogènes, si l'effet est drastique, d'autres VSI entrainent des modifications en phase. Notamment, un nombre important de mutations a été identifié, au sein de notre laboratoire, comme responsable d'un saut en phase, partiel ou total, de l'exon 3 (Δ3) de *BRCA2* (données non publiées). Les conséquences de ces anomalies d'épissage sur la fonctionnalité de la protéine BRCA2 résultante demeurent inconnues. Afin de mieux comprendre la contribution de la pathogénicité du saut partiel de l'exon 3 de *BRCA2* induit par certaines variations, il serait important d'évaluer l'effet combiné de ces mutations à la fois au niveau de l'épissage de l'ARN et de sur la fonctionnalité de la protéine. L'étude de ce modèle pourrait apporter des pistes contribuant à l'interprétation clinique de VSI conduisant à différents niveaux de Δ3 et plus généralement de VSI conduisant à des défauts d'épissage partiels. (Résultats – Chapitre III).

Partie III : Résultats

CHAPITRE I : EVALUATION A LARGE ECHELLE DE LA FIABILITE DES APPROCHES DE PREDICTION D'ALTERATIONS DES ESR

Le développement du séquençage d'ADN à haut débit et son implémentation, à la fois en recherche et en génétique médicale, représente une avancée majeure pour la détection de variations nucléotidiques potentiellement associées à des maladies génétiques. Ce progrès technologique a permis non seulement la découverte de la variabilité insoupçonnée du génome humain, mais également l'identification d'un nombre, plus important que prévu, de variations nucléotidiques de signification biologique inconnue (VSI), y compris chez des patients évocateurs de pathologies héréditaires. Actuellement, un des obstacles les plus importants en génétique médicale est l'interprétation de ces variations, en particulier les substitutions ponctuelles (SNV, *single nucleotide variant*). Dans ce cadre, une attention particulière est souvent donnée aux variations qui génèrent des changements au niveau protéique, i.e. les mutations faux sens et les non-sens, tout en négligeant leur impact potentiel sur l'épissage de l'ARN.

Il a cependant été démontré, récemment, qu'un nombre inconsidéré de SNV exoniques ont un effet sur l'épissage et en particulier via l'altération des éléments exoniques de régulation d'épissage (ESR). Jusqu'ici, il était admis que, contrairement aux variations qui touchent les sites d'épissage, les effets de celles touchant des ESR potentiels étaient difficiles à prédire. En utilisant des jeux de données dérivés d'études expérimentales, sur des variations localisées dans *MLH1* exon 10, *BRCA2* exon 7, *BRCA1* exon 6, *CFTR* exon 12 et *NF1* exon 37, nous avons récemment démontré, de façon prometteuse, le pouvoir prédictif des deux nouvelles approches bioinformatiques de prédictions d'altération des ESR, QUEPASA et HEXplorer, basées sur les scores ΔtESRseq et ΔHZEI, respectivement. Néanmoins, il est nécessaire, avant d'inclure ces approches dans les stratégies de stratification de variations exoniques pour des tests fonctionnels d'épissage, de les évaluer à plus large échelle.

Par conséquent, nous avons entrepris une analyse comparative à grande échelle de quatre nouvelles approches bioinformatiques visant à prédire les altérations d'ESR potentiels, incluant QUEPASA et HEXplorer. La stratégie utilisée a été de confronter les prédictions bioinformatiques

générées par les quatre approches avec les données expérimentales extraites de la littérature pour plus de 1200 variations. Nos résultats ont confirmé le pouvoir prédictif des approches QUEPASA et HExplorer et ont révélé la fiabilité des approches SPANR et HAL en démontrant que l'ensemble de ces approches étaient capables de prédire, avec de bonnes sensibilité et spécificité : (i) quelles variations exoniques sont les plus susceptibles d'impacter des ESR, (ii) la direction des défauts d'épissage provoqués, ainsi que (iii) la sévérité de ces anomalies.

Dans un second temps, nous avons ensuite entrepris d'augmenter les performances des analyses bioinformatiques dédiées aux ESR (i) en optimisant le pouvoir prédictif de chacune des nouvelles approches et (ii) en combinant, ensemble et de toutes les façons possibles, ces outils bioinformatiques. Nos résultats ont permis de définir des seuils de décision optimaux plus précis pour l'utilisation de ces approches et d'établir que la combinaison de ces outils permet d'améliorer le pouvoir prédictif des analyses dédiées aux ESR, avant d'être validés par des études prospectives sur de nouveaux exons modèles.

En somme, l'ensemble de nos données suggèrent qu'il est possible de prédire les altérations d'éléments exoniques régulateurs de l'épissage par des nouvelles méthodes bioinformatique. Ces nouvelles approches pourront être intégrées, avec les outils de prédictions des altérations des sites d'épissage, dans les stratégies de stratification de variations pour les analyses fonctionnelles d'épissage et faciliter l'identification de mutations potentiellement délétères parmi les nombreuses variations nucléotidiques détectées par NGS, avec des implications pour toutes les maladies d'origine génétique.

Ces travaux font l'objet d'une publication qui sera prochainement soumise à *Nucleic Acid Research* (IF 11.561).

Demystifying the splicing code:
new bioinformatics insights for the interpretation of genetic variants

Hélène Tubeuf [1,2], Camille Charbonnier [1], Omar Soukarieh [1], André Blavier [2], Arnaud Lefebvre[3], Hélène Dauchel[3], Thierry Frebourg[1,4], Pascaline Gaildrat [1], Alexandra Martins [1, §]

1 Inserm-U1245, UNIROUEN, Normandie University, Normandy Centre for Genomic and Personalized Medicine, Rouen, France. **2** Interactive Biosftware, Rouen, France. **3** LITIS EA 4108 Computer Science, Information Processing and Systems Laboratory, UNIROUEN, Normandie University, Mont-Saint-Aignan, France. **4** Department of Genetics, University Hospital, Normandy Centre for Genomic and Personalized Medicine, Rouen, France.

Author contributions

HT, PG and AM conceived and designed the project. HT performed the bioinformatics analyses, generated the experimental data and achieved statistical analyses helped by CC. HT, CC, PG and AM were involved in data interpretation. HT, CC, PG and AM wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Farah Ouechtati for helping to collect the pEx-SRE dataset, and Gaia Castelain for technical assistance. We are grateful to Karim Labrèche, Vivien Deshaies, and Alan Lahure for their participation in the preparation of the HExoSplice interface.

Fundings

This project was supported by the OpenHealth Institute, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (Gefluc) as well as the European Union and Région Normandie. Europe gets involved in Normandie with European Regional Development Fund (ERDF). HT was funded by a CIFRE PhD fellowship (#2015/0335) from the French Association Nationale de la Recherche et de la Technologie in the context of public-private partnership between INSERM and Interactive Biosoftware.

Abstract

With the implementation of high-throughput sequencing in medical genetics, the challenge is no longer the detection of nucleotide changes in patients' DNA but their biological and clinical interpretation. In theory, any intragenic variant can potentially affect RNA splicing by either altering splice sites or splicing regulatory elements (SREs). Yet, it is often presumed that SRE alterations are rare and difficult to predict. Recently, a pilot study showed that a larger-than-expected fraction of exonic variants induces splicing defects by affecting SREs, and that most of these alterations could be predicted by the QUEPASA and HEXplorer SRE-dedicated bioinformatics approaches. Here we report the first large-scale comparative analysis of four SRE-dedicated methods (QUEPASA, HEXplorer, SPANR and HAL) by assessing their performance in predicting the impact on splicing of >1300 exonic variants from >80 disease-causing genes. Our results revealed that the four methods display good predictive power when using decision thresholds derived from ROC curve analyses, with QUEPASA and HAL having the best performances either as stand-alone methods or in combination. This study highlights the potential of SRE-dedicated computational approaches as filtering tools for identifying disease-causing candidates among the plethora of variants detected by high-throughput sequencing and provides guidance on how to use them.

Introduction

Over the past ~fifteen years, the tremendous progress made in high-throughput sequencing technologies has greatly accelerated the detection of nucleotide changes in individual genomes in a time- and cost-effective manner (Goodwin *et al.*, 2016). Massive parallel DNA sequencing is now widely implemented both in routine diagnostics and in basic research, resulting in an expanding archive of human genetic variation (Rabbani *et al.*, 2014; Shendure, 2011). However, the terrific improvement made in variant discovery has outperformed our capacity to interpret most of the detected variations, their biological impact remaining essentially unknown. Accordingly, the clinical interpretation of variants of unknown significance (VUS) has since been recognized as one of the major bottlenecks in current medical genetics (Cooper and Shendure, 2011; Frebourg, 2014).

An important hope emerged with the recent development of four bioinformatics approaches specifically dedicated to splicing signals and potentially suitable to pinpointing variant-induced SRE-modifications: QUEPASA (Di Giacomo *et al.*, 2013; Ke *et al.*, 2011), HEXplorer (Erkelenz *et al.*, 2014), SPANR (Xiong *et al.*, 2015) and HAL (Rosenberg *et al.*, 2015). QUEPASA is based on ESRseq scores that were determined experimentally by assessing the splicing regulatory properties (positive or negative) of all possible hexamers in the context of splicing reporter minigenes containing cassette exons (Ke *et al.*, 2011). In the case of QUEPASA, the effect on splicing of any nucleotide variant can be predicted by calculating ΔtESRseq scores, i.e. the change in total ESRseq scores between variant and WT sequences (Di Giacomo *et al.*, 2013). To our knowledge, no ΔtESRseq calculation tool has yet been made available to the public. HEXplorer is based on a RESCUE-type statistical analysis that computed the relative distribution of hexamer motifs in exons and in introns, the effect of nucleotide variants being inferred from ΔHZ_{EI} score changes calculated by using a tool provided by the authors (Erkelenz *et al.*, 2014). SPANR is a state-of-the-art computational splicing predictor based on a machine learning approach, which was trained on exon skipping events from RNA-seq data and 1393 carefully selected DNA sequence features, the impact of SNVs being predicted by Δψ values that can be obtained by using an openly accessible online interface (Xiong *et al.*, 2015). Finally, HAL modelled in a quantitative manner the contribution of randomised K-mer sequences to the alternative splicing pattern of a synthetic two-exon minigene, the predicted effect of exonic variants being equally based on Δψ score changes that can be retrieved from a website provided by the authors or, eventually, from using a

code that was made publically available (Rosenberg *et al.*, 2015). In a previous study, we performed a small-scale comparative analysis of the three first methods and concluded that QUEPASA and HEXplorer were effective in predicting ESR mutations but, unexpectedly, not SPANR (Soukarieh *et al.*, 2016).

An important hope emerged with the recent development of four bioinformatics approaches specifically dedicated to splicing signals and potentially suitable to pinpointing variant-induced SRE-modifications: QUEPASA (Di Giacomo *et al.*, 2013a; Ke *et al.*, 2011), HEXplorer (Erkelenz *et al.*, 2014), SPANR (Xiong *et al.*, 2015a) and HAL (Rosenberg *et al.*, 2015). QUEPASA is based on ESRseq scores that were determined experimentally by assessing the splicing regulatory properties (positive or negative) of all possible hexamers in the context of splicing reporter minigenes containing cassette exons (Ke *et al.*, 2011). In the case of QUEPASA, the effect on splicing of any nucleotide variant can be predicted by calculating ΔtESRseq scores, i.e. the change in total ESRseq scores between variant and WT sequences (Di Giacomo *et al.*, 2013a). To our knowledge, no ΔtESRseq calculation tool has yet been made available to the public. HEXplorer is based on a RESCUE-type statistical analysis that computed the relative distribution of hexamer motifs in exons and in introns, the effect of nucleotide variants being inferred from ΔHZ_{EI} score changes calculated by using a tool provided by the authors (Erkelenz *et al.*, 2014). SPANR is a state-of-the-art computational splicing predictor based on a machine learning approach, which was trained on exon skipping events from RNA-seq data and 1393 carefully selected DNA sequence features, the impact of SNVs being predicted by Δψ values that can be obtained by using an openly accessible online interface (Xiong *et al.*, 2015b). Finally, HAL modelled in a quantitative manner the contribution of randomised K-mer sequences to the alternative splicing pattern of a synthetic two-exon minigene, the predicted effect of exonic variants being equally based on Δψ score changes that can be retrieved from a website provided by the authors or, eventually, from using a code that was made publically available (Rosenberg *et al.*, 2015). In a previous study, we performed a small-scale comparative analysis of the three first methods and concluded that QUEPASA and HEXplorer were effective in predicting ESR mutations but, unexpectedly, not SPANR (Soukarieh *et al.*, 2016).

Here, we re-evaluated SPANR and assessed the performance of the four SRE-dedicated predictors by performing a comparative analysis using a large training dataset of more than 1200 disease-

associated exonic variants for which there was RNA data available. All methods displayed good predictive power, with HAL and QUEPASA generally showing the best performances either as stand-alone methods or in combination. We then validated our findings with a prospective benchmark dataset of 150 variants. To our knowledge, this is the first study where these four methods are compared side-by-side, providing detailed information on their performance and clues on the best way to use them (e.g. optimized thresholds).

Material & Methods

Splice site dedicated *in silico* analyses. In order to predict variant-induced alterations of reference 3' and 5' splice sites, we resorted to MaxEntScan (MES), and SpliceSiteFinder-like (SSFL) algorithms by following the recommendations of Houdayer *et al.* 2012 (Houdayer *et al.*, 2012) except that here both algorithms were interrogated via the integrated software tool Alamut Batch version 1.5.2 (Interactive Biosoftware, Rouen, France). More precisely, MES and SSFL score changes (Δ) between variant and WT sequences were taken as proxies for the probability of a direct splice site alteration. We considered that a variant was susceptible of negatively affecting a reference or an alternative splice site if ΔMES ≤ -15% and ΔSSFL ≤-5% (Houdayer *et al.*, 2012). Moreover, we considered that a variant was susceptible of creating a competing splice site if local MES scores were equal to or greater than those of the corresponding reference splice site.

SRE-dedicated *in silico* prediction approaches. For the prediction of variant-induced ESR alterations, we resorted to four recently developed SRE-dedicated *in silico* approaches: (i) QUEPASA, which is based in the calculation of total ESR_{seq} score changes ($\Delta tESR_{seq}$) (Di Giacomo *et al.*, 2013a; Ke *et al.*, 2011), (ii) the HEXplorer method which calculates ΔHZ_{EI} values (Erkelenz *et al.*, 2014), (iii) the SPANR approach that provides ΔΨ scores (Xiong *et al.*, 2015a), and (iv) HAL which is also based on the calculation of ΔΨ scores (Rosenberg *et al.*, 2015). The main features of the four approaches are summarized in Supplementary Table S1. Both ΔtESRseq and ΔHZ_{EI} scores were calculated with the Alamut Batch prototype tool version 1.5.2 (ESRseq), developed in collaboration with the Interactive Biosoftware Company (http://www.interactive-biosoftware.com). In addition, we created HExoSplice, a web tool for calculating ΔtESRseq scores, freely accessible at http://bioinfo.univ-rouen.fr/HExoSplice_submit. SPANR and HAL ΔΨ scores

were retrieved from the corresponding online interfaces (http://tools.genes.toronto.edu and http://splicing.cs.washington.edu/SE, respectively). The WT Ψ values (level of WT exon inclusion) entered into the HAL interface were those indicated in Supplementary Tables S3-S5. For each SRE-dedicated *in silico* tool, score changes (Δ) of the exonic variants located outside the splice sites, smaller than the pre-established negative thresholds, as indicated, were considered predictive of increased exon skipping, whereas those higher than the positive thresholds were considered as potentially causing an increase in exon inclusion. The thresholds were sequentially optimised throughout the study. Multi-SRE approaches resulting from different combinations of the stand-alone methods are described below, under Performance assessment of stand-alone and combined SRE-dedicated algorithms**.**

Datasets. Five datasets were used in this study (Supplementary Figure S1), including one dataset containing both exonic and intronic variants (MMR dataset), three datasets exclusively containing exonic variants (pilot, training and validation datasets), and one dataset of intronic substitutions only (preliminary pEx-SRE dataset), as indicated below.

MMR dataset was recently used by Xiong and co-workers to validate SPANR-based predictions (Xiong *et al.*, 2015). This dataset includes both exonic and intronic *MLH1* and *MSH2* (MMR genes) variants for which there is available RNA data, notably 146 variants in *MLH1* and 79 variants in *MSH2*, thus yielding to a total of 225 variants (Supplementary Table S2). Among these variants, 134 are exonic and 91 are intronic. Importantly, 117 (52%) are known as splicing mutations whereas 108 (48%) have no impact on splicing.

Pilot dataset was recently reported by our group (Soukarieh *et al.*, 2016). It exclusively contains exonic variants (n=154), which are distributed within 5 exons from 5 different genes (Supplementary Table S3). The pilot dataset includes variants in *MLH1* exon 10 (n=15), *BRCA2* exon 7 (n=32), *BRCA1* exon 6 (n=42), *CFTR* exon 12 (n=41) and *NF1* exon 45 (n=24). These variants are located outside the 3' and 5' splice sites (i.e. outside the very first or the last 3 positions of the exon) and their effects on splicing are known. Among these variants, 61 (40%) caused splicing defects (50 increased exon skipping, and 11 increased exon inclusion) and 93 (60%) had no impact on splicing.

Training dataset represents a largely extended version of the pilot dataset. It was prepared by adding newly selected variants to the previous collection, thus yielding a training dataset of 1214 exonic variants suitable for SRE-dedicated large-scale *in silico* analyses (and Supplementary Table S4). We prepared the training dataset as follows. First, a large number of exonic variants different from the pilot dataset (n~ 1250, data not shown) were collected from the literature based on the availability of splicing data (RT-PCR results obtained with patients' RNA and/or minigene reporter assays). Then, we retained for further analysis only those located outside the reference 3' and 5'ss, and having no direct impact on the creation of new splice sites according to the available experimental data and to MES *in silico* predictions (n=1060). Finally, these variants were added to the pilot dataset leading to the creation of a training dataset of 1214 exonic variants, distributed within 190 different exons from 88 different genes (Supplementary Figure S1 and Supplementary Table S4).

Validation dataset exclusively contained exonic variants (n=150), which are distributed within 3 exons from 3 different genes and were used in a prospective manner (Supplementary Table S5). More precisely, these subsets encompassed variants in (i) *MSH2* exon 5 (n=22) including 21 variants for which there was no available splicing data and 1 variant previously known to induce exon skipping, (ii) *BRCA1* exon 5 (n=98), including 94variants for which there was no available splicing data and 2 variants previously known to induce exon skipping and (iii) *MAPT* exon 10 (n=30), including 21 variants for which there was no available splicing data, and 9 variants known to either enhance exon skipping (n=1) or exon inclusion (n=8). The three validation subsets were prepared as follows. First, we retrieved all *MSH2* exon 5, *BRCA1* exon 5 and *MAPT* exon 10 variants reported in human variation databases (Supplementary Tables S6 to S8). We next excluded those more likely to directly affect the strength of the reference or alternative 3' or 5' splice sites as determined bioinformatically by using MES and SSFL. Whereas all the remaining *BRCA1* exon 5 and *MAPT* exon 10 variants were retained for functional analyses (n=98 and n=30, respectively), the *MSH2* exon 5 subset was narrowed to a group of 22 variants half of which were predicted as the most susceptible to induce exon skipping (n=11) and the other half the less likely to do so (n=11) according to the SRE-dedicated methods (QUEPASA, HEXplorer SPANR and HAL), as described in Supplementary Figure S2. Finally, we determined the impact on RNA splicing of the 150 variants from this dataset by performing minigene splicing assays as described below.

Preliminary pseudoexon-SRE (pEx-SRE) dataset. This dataset contains deep intronic substitutions only (n= 13), which were gathered by querying the literature for variants responsible for pseudoexon (pEx) creation due to the potential alteration of deep intronic SREs (i.e. variants inducing local pEx inclusion but not directly affecting pEx splice sites) (Supplementary Figure S8). The pEx-SRE dataset was prepared as follows. We first performed bibliographic searches by using the keywords "pseudoexon" (or "cryptic exon"), "mutation" (or "variation") and "intronic" yielding a total of ~160 pEx-inducing variants distributed within ~80 different genes (data not shown). Most of these variants created new 3' or 5' splice sites that contributed to the activation of a nearby cryptic 5'ss or 3'ss, respectively, leading to the exonisation of the intervening sequence. From this collection, we then extracted all variants identified within the pEx sequences but located outside of the corresponding splice sites (n=13).

Performance assessment of stand-alone and combined SRE-dedicated algorithms. We attempted to improve SRE-dedicated *in silico* predictions by combining in every possible way the outcomes of the four SRE-dedicated bioinformatics approaches (QUEPASA, HEXplorer SPANR and HAL). We took into account two scenarios: variant-increased exon skipping and variant-induced exon inclusion. Our decision rules were based on the following options. First, each approach was tested in a stand-alone mode taking into account the optimal thresholds estimated from Receiver Operating Characteristic (ROC) curve analysis by minimizing the distance at the top left corner ("closest to top left" criteria, i.e. by maximizing both specificity and sensitivity). Then, we tested combinations of multiple SRE approaches either as duo, trio or tetra options as follows: (i) each pair of approaches in either a "both" or "either" mode, (ii) "at least two" out of the four approaches, (iii) each trio of approaches in either an "all" or "either" mode, (iv) "at least three" out of the 4 approaches, (v) all four approaches, (vi) "at least one" out of the four approaches, and finally, (vii) a linear regression taking into account all four approaches, for predicting either variant-induced exon skipping ($LR_{skip} = \left(\frac{e^{-1,451752 - (0,518937\times\Delta tESRseq) - (0,009326\times\Delta HZEI) - (0,030534\times\Delta\Psi SPANR) - (0,021497\times\Delta\Psi HAL)}}{1+ e^{-1,451752 - (0,518937\times\Delta tESRseq) - (0,009326\times\Delta HZEI) - (0,030534\times\Delta\Psi SPANR) - (0,021497\times\Delta\Psi HAL)}}\right)\times 100$) or variant-increased exon inclusion ($LR_{inc} = \left(\frac{e^{-2,328067 - (0,427250\times\Delta tESRseq) + (0,004037\times\Delta HZEI) - (0,041274\times\Delta\Psi SPANR) - (0,034426\times\Delta\Psi HAL)}}{1+ e^{-2,328067 - (0,427250\times\Delta tESRseq) + (0,004037\times\Delta HZEI) - (0,041274\times\Delta\Psi SPANR) - (0,034426\times\Delta\Psi HAL)}}\right)\times 100$). All coefficients taking into account in linear regressions were significant (two-sided p-value < 0.01), except for the HEXplorer inclusion coefficient (two-sided p-value = 0.2463). The performances of the different

decision rules were then evaluated by jackknife cross-validation i.e. by leaving one variant out of the training dataset, recalibrating the model and predicting the observation that was left out. This operation was repeated until each variant has played the role of a validation sample. Then, we selected the three best multi-SRE decision rules and compared their predictive power to the stand-alone methods by measuring six parameters: sensitivity, specificity, accuracy, positive predictive value, negative predictive value and Matthew's correlation coefficient. These parameters were defined as follows: sensitivity (Sen) = [TPx100/(TP + FN)], specificity (Spe) : [TNx100/(TN + FP)], accuracy (Acc) = [(TN+TP)x100/(TN+TP+FN+FP), positive predictive value (PPV) : [TPx100/(TP+FP)], negative predictive value (NPV): [TNx100/(TN+FN)] and Matthew's correlation coefficient (Mcc) : [(TPxTN-FPxFN)/√((TP+FN)x(TP+FP)x(TN+FN)x(TN+FP))], where TP (true positive) and FN (false negative) are the numbers of positive samples that are predicted/called to be positive and negative, respectively. Analogously, TN (true negative) and FP (false positive) are the numbers of negative samples that are predicted to be negative and positive respectively. TP, TN, FP, FN were determined by taking into account thresholds, as indicated, estimated from ROC curve analysis by minimizing the distance at the top left corner. ROC curves were performed by using GraphPad Prism software (Version 5.0) and easyROC (http://www.biosoft.hacettepe.edu.tr/easyROC/). The predictive power of all bioinformatics tools were then visually compared to each other by using Venn diagrams plotted by Jvenn (Bardou *et al.*, 2014), an interactive web application (http://jvenn.toulouse.inra.fr/app/example.html).

Minigene splicing assays. In order to evaluate the impact of the selected *MSH2* exon 5, *BRCA1* exon 5 and *MAPT* exon 10 variants on RNA splicing, we performed functional assays based on the comparative analysis of the splicing pattern of wild-type (WT) and mutant reporter minigenes, as follows. Minigenes were prepared by using two different vectors: pSPL3mK and pCAS2 (Supplementary Figure S3). The wild-type genomic fragments containing the exons of interest and at least 150 bp of their flanking intronic sequences, i.e. *MSH2* [c.793-228_c.942+186], *BRCA1* [c.135-153_c.212+175] and *MAPT* [c.823-181_915+182], were inserted into the BamHI and MluI cloning sites of the reporter plasmids pCAS2 or pSPL3mK, yielding the three-exon hybrid minigenes pCAS2-*MSH2*e5, pCAS2-*BRCA1*e5 and pSPL3mK-*MAPT*e10, respectively. Minigenes carrying nucleotide variants in *MSH2*e5, *BRCA1*e5 or *MAPT*e10 were prepared by site-directed mutagenesis by using the two-stage overlap extension PCR method (Ho *et al.*, 1989), a combination of specific primers indicated in Supplementary Table S9 and the WT constructs as template. Then,

the mutant amplicons were introduced into a previously linearized vector at BamHI and MluI cloning sites by homologous recombination using the SLICE method (Motohashi, 2015). All constructs were sequenced to ensure that no unwanted mutations had been introduced into the inserted fragments during the PCR or cloning process. Next, WT and mutant minigenes (400 ng/well) were transfected in parallel into HeLa cells grown at ~70% confluence in 12-well plates using the FuGENE 6 transfection reagent (Roche Applied Science). HeLa cells obtained from ATCC were cultivated in Dulbecco's modified Eagle medium (Life Technologies) supplemented with 10% fetal calf serum in a 5% CO2 atmosphere at 37°C. Twenty-four hours later, total RNA was extracted using the NucleoSpin RNA II kit (Macherey Nagel) according to the manufacturer's instructions, and the minigenes' transcripts were analysed by semi-quantitative RT-PCR (30 cycles of amplification) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng total RNA and minigene specific primers (Supplementary Table S9). RT-PCR products were separated by electrophoresis on 2.5% agarose gels containing ethidium bromide and visualized by exposure to ultraviolet light under saturating conditions using the Gel Doc XR image acquisition system (Bio-RAD), followed by gel-purification and sanger sequencing for proper identification of the minigene's transcripts. Finally, splicing events were quantitated by performing equivalent fluorescent RT-PCR reactions followed by capillary electrophoresis on an automated sequencer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) and computational analysis by using the GeneMapper v5.0 software (Applied Biosystems). Results are presented as the mean of three independent experiments.

Statistical analyses. Data derived from confrontation of experimental and *in silico* analyses were compared by using either Student's test, one-way ANOVA test and Pearson's correlation coefficient or their derivatives depending on data distribution patterns as detailed in Supplementary Table S10. Broadly, Mann-Whitney (non-Gaussian distribution) or Student's test were used for assessing the performance of the bioinformatics tools when only 2 groups of variants were taken into account (i.e. variants that induced exon skipping versus those that did not). Similarly, the Kruskal-Wallis or ANOVA tests followed by either Duns or Bonferroni post-tests, respectively, were used for assessing the performance of the bioinformatics tools in discriminating 3 groups of variants (i.e. variants that increased exon skipping versus those with no effect on splicing versus those that increased exon inclusion). Linear correlation between exon inclusion levels and *in silico* predictions was measured by calculating Spearman or Pearson correlation coefficients. All

statistical analysis were performed by using GraphPad Prism software (Version 5.0). Results are expressed as two sided p-values (* p-value<0.05, ** p-value<0.01, *** p-value<0.001) and were considered significant when p-value <0.05).

Databases. BIC (Breast Cancer Information Core, https://research/nhgri.nih.gov/bic/) (Szabo *et al.*, 2000), BRCA ShareTM (Béroud *et al.*, 2016; Caputo *et al.*, 2012), dbSNP (the Single Nucleotide Polymorphism database http://www.ncbi.nlm.nih.gov/SNP/), COSMIC (Catalogue of Somatic Mutations in Cancer, http://cancer.sanger.ac.uk/cosmic) (Forbes *et al.*, 2017), ClinVar (https://www.ncbi.nlm. nih.gov/clinvar/) (Landrum *et al.*, 2016), ExAC (Exome Aggregation Consortium, http://exac.broadinstitute.org/), ESP (Exome Sequencing Project, http://evs.gs.washington.edu/EVS/), HGMD (Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/index.php) (Stenson *et al.*, 2017), UMD-*MSH2* (Universal Mutation Database-*MSH2*, http://www.umd.be/*MSH2*/) (Grandval *et al.*, 2013), LOVD (Leiden Open Variation Database, http://chromium.lovd.nl/LOVD2/colon_cancer/home.php? select_db=*MSH2* ; http://chromium.lovd.nl/LOVD2/cancer/home.php?select_db=BRCA2) (Fokkema *et al.*, 2011; Plazzer *et al.*, 2013), ALZFORUM (http://www. alzforum.org/genetics).

Nomenclature. The identity of the genetic variants is described by following the nomenclature guidelines of the Human Genome Variation Society (HGVS) and taking into account the reference sequences of each gene as indicated in Supplementary Tables S2-S4. Nucleotide numbering is based on the cDNA sequence, c.1 denoting the first nucleotide of the translation initiation codon.

Results

Besides QUEPASA and HEXplorer, SPANR and HAL are also promising tools for predicting ESR mutations.

Recently, we reported that splicing mutations, especially those affecting potential ESRs, were more prevalent than initially expected, and that these could now be bioinformatically predicted by QUEPASA and HEXplorer but, unexpectedly, not by the SPANR approach (Soukarieh *et al.*, 2016). More concretely, QUEPASA and HEXplorer showed good sensitivity and specificity in predicting variant-increased exon skipping at the score thresholds of -0.5 and -20, respectively,

whereas SPANR had low-to-null sensitivity at the recommended threshold of -5% (-0.05) (Soukarieh *et al.*, 2016). This conclusion stemmed from the analysis of a pilot dataset of 154 exonic variants located within five different exons at positions non-overlapping splice sites, with a relatively balanced number of variants causing an impact on splicing as compared to those with no effect. Two main questions remained: (i) what could be the reason(s) for SPANR's poor performance? and (ii) could SRE-dedicated *in silico* approaches predict ESR-mutations in other exons/genes? Before engaging in a large-scale analysis, we thus re-assessed the performance of SPANR both by inspecting the MMR dataset initially used by Xiong and co-workers to validate this method (Xiong *et al.*, 2015b) and by revisiting our pilot dataset (Soukarieh *et al.*, 2016). As shown in Figure 1A and Table S2, we noticed that the MMR dataset was particularly enriched in variations mapping at the 3' and 5'splice sites (105 out of 225 variants i.e. 47%), and contained very few splicing mutations outside these positions. In fact, only 17 out of the 225 variants (7%) were putative ESR-disrupting mutations. Our ROC curve analysis for correlating sensitivity and specificity revealed a very high value for the area under the curve for SPANR (AUC=0.91) when taking into account the full set of MMR variants, which is indicative of an excellent predictive power (Figure 1B). However, when the variants were split into two groups (intronic versus exonic), the ROC curve profiles became very different. The SPANR predictions for the intronic variants were extremely good (AUC=0.99) but much less pertinent for discriminating splicing mutations located at exonic positions. Indeed, when the exonic variants were analysed separately from those at intronic positions, the AUC decreased to 0.76. Nevertheless, although lower than for the intronic variations, the AUC values for the exonic subset were still reasonable and suggestive of a relatively good predictive performance. We thus suspected that the threshold of -5% previously established by Xiong *et al.* (seemingly calibrated by the authors to guarantee ≥95% specificity) was appropriate for predicting the impact on splicing of variants located within splice sites, but too stringent for those located at exonic positions potentially affecting ESRs. We then performed a statistical analysis on the pilot dataset used in our previous study (Soukarieh *et al.*, 2016). This time, in order to achieve higher statistical power, we took into consideration the full number of data points (154 exonic variants altogether, Supplementary Table S5), instead of separated into 5 exon-specific subsets as done previously. In these conditions, and contrary to our initial appreciation, we observed that SPANR, similarly to QUEPASA and HEXplorer, and now also HAL, was able to significantly distinguish the exonic variants that affected RNA splicing from those that did not (t-

test or its derivatives, p-value <0.001) and also to predict the severity of the induced splicing defects (Pearson or Spearman correlation as indicated, p-value <0.001), i.e. the score values correlated with the levels of exon skipping (Figure 1C). Our new results thus suggest that SPANR and HAL are additional promising tools for pinpointing ESR mutations. ROC curves further revealed a good predictive performance for SPANR (AUC of 0.78) and confirmed better predictive powers for QUEPASA, HEXplorer and HAL with AUCs of 0.87, 0.86 and 0.87, respectively (Figure 1D). A more detailed view of our findings in terms of sensitivity, specificity and accuracy, at given thresholds, is shown in Figure 1E. One instructive finding from the ROC curves established with the pilot dataset was the observation that one was able to improve SPANR sensitivity (while still keeping good specificity) by decreasing the threshold by ~ one order of magnitude from -5% to -0.38%. This adjustment in the SPANR threshold allowed for a dramatic improvement in sensitivity (4-fold increment, from 18% to 76%) with a moderate impairment in specificity (decrease from 97 % to 74%), thus seemingly rehabilitating SPANR for predicting exon-skipping ESR-mutations. Therefore, we decided to perform a large-scale analysis with all four SRE-dedicated approaches, as they all seemed promising for pinpointing exon-skipping mutations potentially affecting ESRs.

Large-scale analyses show that the four SRE-dedicated approaches can retrospectively predict variant-increased exon skipping

To better evaluate the four SRE-dedicated approaches, we performed a large-scale comparative analysis by extending our initial study from 154 to 1214 exonic variants. The 1060 variants added to the pilot dataset were retrieved from the literature upon searching for exonic variants located outside splice sites and for which there was available RNA data derived either from minigene splicing assays and/or from the analysis of patients' RNA. As shown in Supplementary Figure S1B and Table S4, this large training dataset contains 22 subsets of variants mapping to 190 different exons and 88 different genes. One can distinguish two main groups within the overall collection: a group of variants causing splicing defects (533/1214, i.e. 44%) and a group with no effect on splicing (681/1214, i.e. 56 %) (Supplementary Table S4). The majority of the splicing mutations induced exon skipping (385/533, i.e. 72 %) whereas a small fraction increased exon inclusion (151/533, i.e. 28%). Accordingly, we separated the training dataset into three groups corresponding

to different splicing effect directions as follows: no effect on splicing (681/1214, i.e. 56 %), increased exon skipping (382/1214, i.e. 32%) and increased exon inclusion (151/1214, i.e. 12%).

Statistical analysis performed with this large dataset confirmed that the four approaches were able to (i) discriminate variants that increased exon skipping from those that do not (t-test or its derivatives, p-values <0.001) and (ii) predict direction and severity of the splicing defects (ANOVA, p-values <0.001, and Pearson or Spearman correlation as indicated, p-values <0.001, respectively) (Figures 2B, S4 and S5). ROC curve analysis further confirmed that the four tools displayed good predictive power with better AUC values for QUEPASA, HEXplorer and HAL (0.82, 0.79 and 0.80, respectively) than for SPANR (0.73) (Figure 2C). We then examined the sensitivity, specificity and accuracy of the four methods in predicting variant-induced exon skipping. First, optimal decision thresholds were determined from ROC curve analysis yielding the following optimal cut-off values corresponding to the best balance between sensitivity and specificity: ≤-0.5 for QUEPASA, ≤-14 for HEXplorer, ≤-0.1% for SPANR and ≤-3.4% for HAL. Except for QUEPASA, these values were slightly less stringent than the ones previously determined with the pilot dataset. Next, variants with score differences smaller than the newly established thresholds were considered as the most susceptible to cause exon skipping, allowing to estimate the relative number of true and false calls (TC and FC) generated by each approach (Figure 2D). In these conditions, QUEPASA and HAL produced the highest number of true calls (TC=919 and TC=899 i.e. 76% and 74%, respectively), outperforming HEXplorer and SPANR (TC=860 and TC=813, i.e. 71% and 69%, respectively) in discriminating variants that induce exon skipping from those that do not. Overall, we found the following ranking in sensitivity, specificity and accuracy in predicting exon skipping: QUEPASA > HAL > HEXplorer > SPANR (Figure 2D). Moreover, calculation of Matthews's correlation coefficient (MMC) further indicated a similar trend in performance, notably QUEPASA (0.49) > HAL (0.45) > HEXplorer (0.40) >SPANR (0.36). To evaluate the potential interest of using these approaches as diagnostic decision tools, we then examined positive and negative predictive values (PPV and NPV). We found that the four tools displayed very high NPV for predicting exon skipping (83-87% depending on the approach) but a more moderate PPV (50-59%). We surmise that QUEPASA, HEXplorer, SPANR and HAL are indeed promising as filtering tools for pinpointing ESR-mutations that induce exon skipping, especially because they allow to exclude variants less likely to do so while retaining those that should be experimentally tested in priority. Importantly, by increasing statistical power in the

training dataset relative to the pilot dataset, we found that all four approaches, including SPANR, can predict both the direction and the severity of the induced splicing defects. We also conclude that QUEPASA and HAL show the best performance in predicting variant-induced exon skipping followed by HEXplorer and SPANR, and that they can be applied to a large number of different exons/genes.

Then, we wondered if we could improve variant-increased exon skipping predictions by using the different approaches in a combined way (multi-SRE approach) instead of in a stand-alone mode. The incomplete overlap between the false calls produced by each approach (Figure 2E), suggested that the different methods offered complementary information, and that their combination could reduce the number of incorrect predictions. We thus tried multiple combinations (several duo, trio and tetra options) and compared their predictive power by using cross-validation to determine the most reliable decision rule (Figure 2F). Our results indicate that the prediction models "QUEPASA&HAL" (duo), "at least 3" (trio) and "Linear Regression-Skip (LR_{skip})" (tetra) displayed the best balance between sensitivity and specificity in predicting exon skipping (63 and 87%; 64 and 85%; and 78 and 75%, respectively; Figure 2F). We then applied these three decision rules to our training dataset and compared their performance with those of the stand-alone approaches at the thresholds inferred from the ROC curve analysis. Of note, based on ROC curve analysis (Figure 2C), the threshold inferred for LR_{skip} was 31.1%, i.e. variants with $LR_{skip} \geq 31.1\%$ were considered as the most susceptible to induced exon skipping. As shown in Figures 2D and 2G, each selected decision rule generated a lower number of false calls than each approach alone. Indeed, the three combinations displayed slightly better accuracies than each stand-alone approach, reflecting concomitant better specificities and slightly decreased sensitivities for QUEPASA&HAL and "at least 3", and slightly better sensitivity for LR_{Skip}. We also observed a better PPV (60-69% versus 50-59%) and a somewhat similar NPV (81-88% versus 83-87%) (Figure 2D). Moreover, MMC was increased when the approaches were combined as compared to the stand-alone methods (0.49-0.52 range versus 0.36-0.49 range) (Figure 2D). We conclude that at the pre-established thresholds the SRE-dedicated approaches can overall predict exon skipping with ≥69% accuracy and that combining the approaches helps improving predictions of this type of splicing mutations (≥77% accuracy).

Large-scale analyses indicate that the four SRE-dedicated approaches can retrospectively predict variant-increased exon inclusion

To further evaluate the predictive potential of the four SRE-dedicated approaches, and given the good predictive power in determining the direction of the observed splicing effects (no effect versus induced exon skipping versus increased exon inclusion, ANOVA, p-value <0.001, Figure 2B), we extended our study to the prediction of variant-increased exon inclusion (Figure 3). Statistical analysis revealed that these approaches can indeed discriminate variants that increased exon inclusion from variants those that did not (t-test or its derivatives, p-values <0.001) (Figure 3B). ROC curve analysis further confirmed that the four methods display good predictive power for estimating increased exon inclusion, their performance ranking as follows QUEPASA = HAL > HEXplorer > SPANR (AUC of 0.78, 0.78, 0.73 and 0.71, respectively) (Figure 3C). Moreover, we inferred the following optimal thresholds for predicting increased exon inclusion: ≥+0.36 for QUEPASA, ≥+9.0 for HEXplorer, ≥+0.3% for SPANR and ≥+1.0% for HAL. Variants with scores greater than these thresholds were thus considered as the most susceptible to increase exon inclusion. We then compared the *in silico* predictions with the splicing data (Supplementary Table S4) in order to calculate the relative number of true and false calls produced by each method. As shown in Figure 3D, we observed that HAL produced the highest number of true calls (TC=902, 74%), outperforming SPANR, QUEPASA, and HEXplorer (TC= 825, 831, and 810 i.e. 70%, 69%, and 67%, respectively) in identifying variant-increased exon inclusion. HAL also displayed better sensitivity, specificity and accuracy in predicting variant-increased exon inclusion as compared to SPANR, QUEPASA, and HEXplorer, in this order (Figure 3D). Calculation of MMC further underlined a similar ranking in predictive power such that HAL (0.35) > SPANR (0.29) > QUEPASA (0.26) > HEXplorer (0.23). By examining positive and negative predictive values, we found that the four *in silico* tools display a very high NPV (93% to 95% depending on the approach) but a modest PPV (22% to 29%) in predicting increased exon inclusion (Figure 3D). A visual comparison of the number and distribution of false calls suggested that these could be reduced by overlapping the different methods (Figure 3E and Supplementary Table S4). Again, this prompted us to combine the four SRE-dedicated approaches in multiple ways and to compare them to the stand-alone versions in an attempt to reduce the overall number of incorrect predictions. Based on ROC curve analysis (Figure 3C), the threshold inferred for LR_{inc} was ≤ 6.2%. As shown in Figure 3F, amongst all the multi-SRE approaches, the "QUEPASA&HAL", "at least 3" and "Linear

Regression-Inc (LR_{inc})" decision rules showed the best compromise between sensitivity and specificity in predicting variant-increased exon inclusion (56% and 84%; 63% and 79%; 72% and 73%, respectively). These three decision rules were then applied to our training dataset (Supplementary Table S4). As shown in Figure 3D, "QUEPASA&HAL" and "at least 3" generated lower numbers of false calls (FC=241 and 283, respectively) than each approach alone ($312 \leq FC \leq 404$), unlike the "LR_{inc}" model (FC=321). "QUEPASA&HAL" and "at least 3" displayed better accuracies (82% and 77%, respectively) than each approach alone (67%-74% range), reflecting better specificities (84% and 79%, respectively, compared to 67%-75% range), but lower sensitivities (56 % and 63% compared to 67%-74% range). We also observed slightly better PPV (30 % and 33%, respectively, compared to 22%-39% range) and persistently high NPV (93% and 94% compared to 93%-95% range). Moreover, "QUEPASA&HAL", "at least 3" and "LR_{inc}" showed somewhat better MCC as compared to each method alone (0.30-0.33 versus 0.23-0.29 ranges) except for HAL which was better as stand-alone (MCC=0.35). In terms of accuracy and compared to HAL (the best stand-alone method in this context), the predictions of increased exon inclusion benefitted from implementing the "QUEPASA&HAL" and "at least 3" decision rules but not the "LR_{inc}" linear regression model.

Overall, our results indicate that combined QUEPASA&HAL" and "at least 3" SRE-dedicated approaches improve the ability of predicting variant-increased exon inclusion if one takes into account the following thresholds: +0.36 for QUEPASA, +9.0 for HEXplorer, +0.3% for SPANR and +1.0% for HAL. It is however apparent that the four tools are more efficient for predicting increased exon skipping than for predicting increased exon inclusion. The major limitation in predicting variant-induced inclusion seems to be the relative high number of false positive calls (Figures 3E and 3G).

SRE-dedicated prediction models can prospectively pinpoint exonic splicing mutations and potentially help molecular diagnostics

To validate our prediction models, we then performed a prospective study by focusing on 150 additional variants identified in molecular diagnostic settings but for which very limited RNA splicing data were available (Supplementary Table S5). Three clinically meaningful exons were

selected to build our validation dataset namely: (i) *MSH2* exon 5 (n=22 selected variants) essentially because total skipping of this exon is known to cause Lynch syndrome (Auclair *et al.*, 2006), (ii) *BRCA1* exon 5 (n=98 selected variants), given that alterations of its alternative splicing pattern are associated with hereditary breast and ovarian cancer (Claes *et al.*, 2002), and (iii) *MAPT* exon 10 (n=30 selected variants) because it represents an emblematic alternatively spliced exon for which alterations in either direction have been implicated in Pick's disease (PiD, caused by increased skipping of exon 10) or in frontotemporal dementia with Parkinsonism linked to chromosome 17 (FTDP-17, caused by an increase in exon 10 inclusion) (Iqbal *et al.*, 2016; Liu and Gong, 2008). To assess the impact on splicing of the full set of validation variants, we performed cell-based splicing assays with pCAS2-*MSH2*e5, pCAS2-*BRCA1*e5 and pSPL3mK-*MAPT*e10-derived minigenes. As shown in Supplementary Table S5: (i) the wild-type pCAS2-*MSH2*e5 minigene generated 99% of transcripts containing *MSH2* exon 5, closely recapitulating the physiological splicing pattern of this exon, which undergoes very low level of alternative splicing (5%) as determined by targeted RNA-seq (Brandão *et al.*, 2019), (ii) the wild-type pCAS2-*BRCA1*e5 minigene generated three types of transcripts, one containing the entire *BRCA1* exon 5 (82%, exon inclusion), another containing exon 5 deleted of its last 22 nt (14%, Δ5q(22nt)) and a minor product lacking exon 5 (4% exon skipping), broadly reproducing the alternative splicing pattern of natural *BRCA1* transcripts (Colombo *et al.*, 2014) and (iii) the wild-type pSPL3mK-*MAPT*e10 minigene predominantly generated two types of transcripts, one containing exon 10 and the other lacking exon 10 (~60% exon skipping). Therefore, the wild-type pSPL3m-*MAPT*e10 minigene also closely mimicks the alternative splicing pattern of natural *MAPT* transcripts, which were estimated to undergo ~50% exon 10 skipping in normal human adult brain, notably based on the fact that this tissue expresses approximately equal levels of 3R-tau and 4R-tau protein (Goedert and Jakes, 1990; Goedert *et al.*, 1989; Kosik *et al.*, 1989).

Functional analysis of the validation dataset in the context of these minigene constructs, revealed that 103 out of the 150 variants (69%) altered the splicing pattern of the exon of interest relative to wild-type and that the variants could be separated into three groups as follows: (i) variants that increased exon skipping (n= 55, i.e. 37%), (ii) variants that did not affect splicing (n= 47, i.e. 31%) and (iii) variants that increased exon inclusion (n= 48, i.e. 32%) (Figure 4A and Supplementary Table S5). More precisely, 12 out of the 22 variants (55%) mapping to *MSH2* exon 5 induced exon skipping, whereas 10 variants showed no effect on splicing. In the case of *BRCA1* exon 5, 62 out

of the 98 variants (63%) caused splicing alterations: 29 increased exon skipping, 33 increased exon inclusion and the remaining 36 showed no effect on splicing. Finally, 29 out of the 30 variants (97%) of *MAPT* exon 10 altered its splicing pattern: 14 variants increased exon skipping and 15 variants increased exon inclusion. Only 1 variant showed no effect on splicing (c.851T>G). These assays have thus confirmed the splicing outcomes of 12 previously reported ESR-mutations (Iqbal *et al.*, 2016; Liu and Gong, 2008; Sanz *et al.*, 2010; Tournier *et al.*, 2008) and identified 91 new splicing mutations. Some of these variants may be pathogenic and will deserve further investigation by clinicians and geneticists. For instance, given its drastic impact on *MSH2* exon 5 splicing, we suspect that c.906G>A, a synonymous *MSH2* variant, may cause Lynch syndrome instead of being likely benign as currently assumed (ClinVar database).

Next, we used the output of the minigene assays to evaluate the performance of the four SRE-dedicated *in silico* approaches and of the three selected multi-SRE models in predicting variant-induced splicing alterations. As shown in Figure 4B and 4C, and depending on the *in silico* tool taken into account, variant-induced exon skipping was correctly predicted for most of the variants by both the stand-alone methods (TC=73-87%) and their combinations (TC=85-88%); and equivalent performances were observed for predictions of variant-increased exon inclusion (TC=75-81% and TC=83-87%, respectively). Nevertheless, the stand-alone approaches displayed lower accuracy in predicting exon skipping (78%, 73% and 81% for HEXplorer, SPANR and HAL, respectively) than "QUEPASA&HAL", "at least 3" and "LR_{Skip}" (88%, 85% and 86%, respectively) except for QUEPASA (87%), which was as powerful as the multi-SRE approaches (Figure 4B). In regards to predictions of increased exon inclusion, we observed a better accuracy for the latter (87%, 85% and 83%, respectively) than for the four stand-alone approaches (75% to 81%) (Figure 4C). Statistical analyses further highlighted the overall good performance of QUEPASA, HEXplorer, HAL, SPANR, LR_{skip} and LR_{inc} for discriminating variants that lead to either exon skipping or exon inclusion (Student's test or derivatives, p-values <0.001) and for predicting the direction and the severity of the splicing defects (ANOVA or its derivatives, and Pearson or Spearman correlation, respectively, p-values <0.001) (Supplementary Figures S6 and S7). We conclude from these findings that *MSH2* exon 5, *BRCA1* exon 5 and *MAPT* exon 10 contain a high number of spliceogenic variants affecting potential ESRs, most of which can be pinpointed by the four SRE-dedicated bioinformatics approaches and/or their combinations. These results validate the conclusions derived from the analysis of the training dataset in regards to both

variant-induced exon skipping and variant-increased exon inclusion, thus confirming the predictive power of the SRE-dedicated *in silico* approaches, QUEPASA&HAL being overall the most accurate predictive model.

The performance of SRE-dedicated predictors may be exon-dependent.

We next asked if the pre-established thresholds of the stand-alone SRE-dedicated approaches could be applied equivalently to individual exons/genes in the training and validation datasets. Close inspection of the predictive power of each tool on the 25 independent subsets revealed heterogeneous performances depending on the subset (Supplementary Tables S11 and S12). For instance, whereas all SRE-dedicated approaches displayed good performances in predicting variant-induced exon skipping in the cases of the *SMN2* exon 7 or *MSH2* subsets (accuracies in the 88%-98% and 74%-90% ranges, respectively), they did not show consistent predictive power when applied to the *BRCA1* exon 6 nor the *DYSF* subsets (accuracies in the 56%-79% and 44%-87% ranges, respectively, the latter with persistent 0% sensitivity) (Supplementary Table S11). While it is possible that the poor performances observed on certain exons/genes are due to a bias in the composition of the subsets, for example an under-representation of variants increasing exon skipping, this seems unlikely since the composition appears very similar between the above mentioned subsets (n=3/23 exon-skipping variants for *DYSF* versus n=4/31 for *MSH2*, n=4/43 for *BRCA1* exon 6 versus n=4/43 for *SMN2* exon 7) (Supplementary Table S4). Instead, we wondered if the differences in performance could, at least in part, be due to the generic thresholds determined from the overall analysis of the training dataset, which may not be optimal for each subset. Therefore, we decided to infer optimal thresholds for the different subsets by performing individual ROC curve analyses. As shown in Supplementary Tables S11-S14, this strategy improved the predictions of both variant-induced exon skipping and increased exon inclusion for several subsets, including for *BRCA1* exon 6 for which the performances became comparable to those of *SMN2* exon 7 (Supplementary Tables S11 and S13). These observations suggest a potential need to adjust the thresholds according to the exon of interest, eventually depending on their sensitivity to ESR mutations. Interestingly, we did not observe an overall increase in accuracy when optimizing the threshold of the *DYSF* subset (accuracies from 44%-87% with 0% sensitivity to 26%-65% with 67% sensitivity), HAL being the only stand-alone method showing compelling predictive power

for this subset (Supplementary Tables S11 and S13). It is possible that the SRE-dedicated bioinformatics approaches may not be sufficiently efficient for the analysis of certain genes, such as DYSF. Alternatively, the DYSF subset may be too heterogeneous for a proper analysis (23 variants distributed within 13 different exons).

QUEPASA and HEXplorer may help predict variant-induced creation of pseudoexons deep within introns.

As SRE-dedicated *in silico* approaches are able to pinpoint variants that increase exon inclusion, we also wondered if these tools could predict variant-induced creation of pseudoexons deep within intronic sequences. These events are still poorly characterized and are typically very difficult to anticipate specially the ones involving alterations in splicing regulation (Dhir and Buratti, 2010; Vaz-Drago *et al.*, 2017). We thus extended the analyses of the SRE-dedicated tools to a preliminary dataset containing 13 variants previously reported in 12 different genes as responsible for the creation of pseudoexons due to an alteration of potential splicing regulators (pEx-SRE dataset, Supplementary Figure S8). This dataset has the disadvantage of being very small and lacking negative cases and thus merely providing a preliminary impression. Because SPANR and HAL cannot query variants deep within introns (Supplementary Table S1), we were limited to only using QUEPASA and HEXplorer in the analysis of the pEx-SRE dataset. As indicated in Supplementary Figure S8, our pEx-SRE dataset includes intronic variants that create pseudoexons with sizes varying between 34 and 345 nucleotides in length, positioned deep within intronic sequences, i.e. far from the splice sites of the nearest natural exons, more specifically from at least 469 to 2053 nucleotides away in our dataset. As for the position of the variants within the pseudoexon sequences, they were found at a distance from the pEx 5' and 3' termini ranging from +9 and -4, to +138 and -235, respectively, i.e. outside of the pEx splice sites. The pEx splice sites were correctly predicted by splice site-dedicated *in silico* approaches (MES and/or SSFL, data not shown) suggesting that these variants induce pseudoexon inclusion by affecting intronic SREs that lead to the activation of pre-existing intronic cryptic 3' and 5' splice sites that would otherwise stay unused by the splicing machinery. By using the threshold values derived from the analysis of the training dataset, we found that 10 out of the 13 pEx-SRE variants were correctly predicted either by QUEPASA or by HEXplorer as causative of pseudoexon inclusion. True-positive predictions

were concordant between the 2 methods for 9 out of these 10 variants. Although very preliminary, our results suggest that SRE-dedicated *in silico* approaches, notably QUEPASA and HEXplorer, may be useful in the future for predicting the creation of variant-induced pseudoexons. To further investigate this hypothesis it will be necessary to create benchmark datasets dedicated to pseudoexon analysis for instance by performing mutagenesis experiments within the pEx-SRE sequences in the context of minigene assays, which will allow to collect data on positive and negative cases that can then be compared to bioinformatics predictions.

Discussion

The present study was initiated to follow up on our observation that ESR-mutations can be predicted by using new SRE-dedicated *in silico* tools, at least in 5 different exons (*MLH1* exon 10, *BRCA2* exon 7, *BRCA1* exon 6, *CFTR* exon 12 and *NF1* exon 37) (Soukarieh *et al.*, 2016). The question remained as to whether these approaches could be equally reliable in identifying ESR-mutations in other exons or genes and be applicable to large-scale studies. To better evaluate the predictive power of the new tools we decided to extend our analysis to ~1200 variants distributed within 88 genes for which functional splicing data were reported in the literature, followed by a prospective study of 150 additional variants identified in 3 clinically-relevant exons. Before this study, only QUEPASA and HEXplorer had been independently validated for the prediction of variant-induced ESR alterations, but not SPANR, which did not show compelling predictive power when using the cut-off value described by Xiong and co-workers (Di Giacomo *et al.*, 2013a; Erkelenz *et al.*, 2014; Grodecká *et al.*, 2017a, 2017b; Ke *et al.*, 2011; Soukarieh *et al.*, 2016; Xiong *et al.*, 2015b). Here we confirmed that QUEPASA and HEXplorer are indeed good predictors and revealed that SPANR can also discern ESR-mutations if one uses an optimized decision threshold (-0.34% instead of -5%). This work further validated HAL, a new bioinformatics approach described as suitable for predicting variant-induced ESR alterations (Rosenberg *et al.*, 2015) that we had not tested before. Overall, our statistical analyses confirmed that QUEPASA, HEXplorer, SPANR and HAL predict, in average, both the direction and the severity of variant-induced splicing defects, underlining their qualitative and quantitative features, and applicability to a large number of exons/genes.

Given the excessive number of exonic variants of unknown significance currently identified in large-scale genetic screenings and the need to prioritize RNA analyses in a time- and cost-effective manner, we inferred optimal decision thresholds for each method by maximizing both sensitivity and specificity in the large-scale analysis, and then implemented three combination approaches (multi-SRE) aiming at further reducing the number of false calls. When taking into account the new thresholds and the full RNA splicing data, both in the training and the validation sets (Figures 2-4), the most accurate stand-alone and multi-SRE approach for predicting induced exon skipping were QUEPASA and QUEPASA&HAL, respectively, whereas the more accurate predictions of increased exon inclusion were produced by HAL and QUEPASA&HAL. Given their consistently better sensitivities (~75%), we suggest to preferentially using QUEPASA and HAL as stand-alone approaches for predicting variant-induced exon skipping and –increased exon inclusion, respectively. However, if the number of experimental analyses are restricted due to limited resources, we recommend applying the combined QUEPASA&HAL method, which has higher specificity than the stand-alone versions at the same thresholds, keeping false positives to a minimum. Alternatively and depending on the purpose of the analysis, users may prefer to apply either high-sensitivity (≥95%) or high-specificity (≥95%) thresholds (Supplementary Figure S9) instead of those chosen in our study, which correspond to the best compromise between these two metrics, and/or to combine the 4 methods as shown in Figures 2F and 3F ("at least 1" or "all4" options). Our recommendations are meant to provide tentative guidance to laboratories wishing to predict ESR-alterations among a large number of exonic variants identified within multiple exons/genes, such as those identified by exome sequencing. QUEPASA and HAL can be easily implemented by using open access online tools (this study and Rosenberg *et al.*, 2015) and have the advantage of being amenable to automation (Ke *et al.*, 2011, Rosenberg *et al.*, 2015). This implies that they can be added to bioinformatics pipelines downstream high-throughput genetic screenings and eventually improve the discovery of disease-causing variants. Of note, the open-access SPANR predictor, as well as the commercial Alamut Batch prototype tool v1.5.2 (used in our work for performing large-scale QUEPASA and HEXplorer predictions) were designed for working with VCF files, such as those generated by next generation sequencing (NGS). It is important to keep in mind though that, according to our data, the current SRE predictors can be useful as filtering tools for stratifying genetic variants for experimental analyses but are not sufficiently robust for directly contributing to clinical decisions. Moreover, the reliability of the

SRE-dedicated methods seems to vary from exon to exon, our results indicating that different cut-off values may be required for optimal performances depending on the exon of interest. For these reasons, if a laboratory wishes to predict the impact on splicing of new variants identified in an exon already evaluated in our work, we suggest using the SRE-dedicated approach that showed the best accuracy with the exon-specific provisory thresholds described in Supplementary Tables S13 and S14.

It is possible that the limited performances observed with certain exons are due to intrinsic exon/gene-specific features that influence their susceptibility to splicing alterations. For example, splicing efficiency can be affected by factors such as: the length of the exon and/or of its flanking introns (Berget, 1995; Fox-Walsh *et al.*, 2005), the strength of 3' and 5'ss (Green, 1991; Shepard *et al.*, 2011), the density in ESRs and in cryptic splice sites (Brillen *et al.*, 2017; Haque *et al.*, 2010; Tammaro *et al.*, 2014), the existence of flanking intronic SREs (Gao *et al.*, 2007; Kashima *et al.*, 2007), as well as the presence of pre-mRNA secondary structures (Buratti and Baralle, 2004; D'Souza and Schellenberg, 2000; Singh *et al.*, 2007). Moreover, splicing patterns can be cell- or tissue-specific due to different expression levels of splicing regulatory factors (Cieply and Carstens, 2015), and may vary depending on nucleosome density and transcription elongation rate (Aissat *et al.*, 2013; Naftelberg *et al.*, 2015). These variables, many of which taken into account by SPANR, reflect the complexity and context-dependence of splicing regulation and underscore the current difficulty of developing highly accurate SRE-dedicated prediction tools. Nonetheless, given the growing knowledge on sequence determinants affecting RNA splicing there is a good possibility that SRE-dedicated *in silico* methods will improve in a near future. For instance, developers of new computational models may benefit from recently established saturation mutagenesis-derived splicing (SMS) scores for >6600 exonic 7-mer sequences (Ke *et al.*, 2018), transcriptome-wide RNA secondary structure maps (Strobel *et al.*, 2018; Sun *et al.*, 2019), extensive outlines of *in vivo* RNA binding sites of splicing regulatory proteins (Park *et al.*, 2016; Yang *et al.*, 2019; Yee *et al.*, 2019), as well as information on splicing-related evolutionary constraints (Savisaar and Hurst, 2018; Wainberg *et al.*, 2016).

Still, we hope that our work will encourage further evaluations of the predictive power of QUEPASA, HEXplorer, SPANR and HAL (and their combinations), notably by testing additional datasets. Such studies may help defining better thresholds for individual exons, as well as bring

clues on strategies to improve SRE predictors. Due to recent technological advances, very large and reliable datasets can now be generated by performing cell-based splicing assays that rely on saturation mutagenesis of either minigene constructs or of genomic DNA, followed by DNA- and RNA-seq for quantifying splicing alterations (Findlay *et al.*, 2018; Ke *et al.*, 2018; Rosenberg *et al.*, 2015). This type of methods, so far applied to very few exons, were instrumental for the development of both QUEPASA and HAL and hold great promise for further improving predictions of variant-induced splicing alterations.

In contrast to saturation mutagenesis, the minigene splicing assays performed in our work (validation dataset) focused exclusively on naturally occurring variants. These experiments uncovered 91 previously unknown splicing mutations in three clinically important exons (11 in *MSH2* exon 5, 62 in *BRCA1* exon 5 and 18 in *MAPT* exon 10) indicating that SRE mutations are still overlooked in molecular diagnostics settings. Most likely *MSH2* exon 5, *BRCA1* exon 5 and *MAPT* exon 10 have a high density in ESRs that help fine-tuning their alternative splicing pattern. Our results thus re-iterate the existence of an important splicing code beyond the protein coding sequence of disease-relevant genes and emphasise the importance of systematically assessing, at least bioinformatically, the potential impact on RNA splicing of variants mapping to these exons. Remarkably, the newly identified splicing mutations include an important fraction of synonymous variants (17 out of 23, i.e. 74% of all synonymous variants in the validation dataset) that either induced exon skipping (n=12) or increased exon inclusion (n=5). What's more, some of the more drastic effects on splicing were observed with synonymous variants (e.g. *MSH2* c.906G>A, *BRCA1* c.165G>A, *MAPT* c.825G>A and *MAPT* c.858C>T, which induced 67%, 87%, 94% and 96% exon skipping, respectively, Supplementary Table S5). Our findings agree with previous observations that a significant proportion of synonymous variants alter RNA splicing (Mueller *et al.*, 2015; Supek *et al.*, 2014), once again highlighting the importance of taking into account the potential impact on splicing of all exonic variants, independently of their coding potential.

In sum, our findings indicate that the SRE-dedicated *in silico* approaches QUEPASA, HEXplorer, SPANR and HAL facilitate the identification of spliceogenic variants in multiple exons/genes suggesting that they can be used in molecular diagnostics settings for prioritizing variants for RNA splicing analyses. It is important to note, however, that clinical classification of spliceogenic variants does not merely depend on experimental RNA results but also in population- and patient-

related data such as variant frequency, clinical phenotype, family history, and co-segregation with disease, among others (de la Hoya *et al.*, 2016; Richards *et al.*, 2015). We hope that our results will be helpful for researchers as well as clinicians, and that they will prompt further evaluations of the performance of SRE predictors as well as strategies for improving their predictive power.

Bibliography

Aissat, A., de Becdelièvre, A., Golmard, L., Vasseur, C., Costa, C., Chaoui, A., Martin, N., Costes, B., Goossens, M., Girodon, E., *et al.* (2013). Combined computational-experimental analyses of CFTR exon strength uncover predictability of exon-skipping level. Hum. Mutat. *34*, 873–881.

Auclair, J., Busine, M.P., Navarro, C., Ruano, E., Montmain, G., Desseigne, F., Saurin, J.C., Lasset, C., Bonadona, V., Giraud, S., *et al.* (2006). Systematic mRNA analysis for the effect of MLH1 and MSH2 missense and silent mutations on aberrant splicing. Hum. Mutat. *27*, 145–154.

Baralle, D., and Buratti, E. (2017). RNA splicing in human disease and in the clinic. Clin. Sci. Lond. Engl. 1979 *131*, 355–368.

Baralle, D., Lucassen, A., and Buratti, E. (2009). Missed threads. The impact of pre-mRNA splicing defects on clinical practice. EMBO Rep. *10*, 810–816.

Baralle, M., Skoko, N., Knezevich, A., De Conti, L., Motti, D., Bhuvanagiri, M., Baralle, D., Buratti, E., and Baralle, F.E. (2006). NF1 mRNA biogenesis: effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. FEBS Lett. *580*, 4449–4456.

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

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. BMC Bioinformatics *15*, 293.

Bauwens, M., De Zaeytijd, J., Weisschuh, N., Kohl, S., Meire, F., Dahan, K., Depasse, F., De Jaegere, S., De Ravel, T., De Rademaeker, M., *et al.* (2015). An augmented ABCA4 screen targeting noncoding regions reveals a deep intronic founder variant in Belgian Stargardt patients. Hum. Mutat. *36*, 39–42.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. *270*, 2411–2414.

Béroud, C., Letovsky, S.I., Braastad, C.D., Caputo, S.M., Beaudoux, O., Bignon, Y.J., Bressac-De Paillerets, B., Bronner, M., Buell, C.M., Collod-Béroud, G., *et al.* (2016). BRCA Share: A Collection of Clinical BRCA Gene Variants. Hum. Mutat. *37*, 1318–1328.

Brandão, R.D., Mensaert, K., López-Perolio, I., Tserpelis, D., Xenakis, M., Lattimore, V., Walker, L.C., Kvist, A., Vega, A., Gutiérrez-Enríquez, S., *et al.* (2019). Targeted RNA-seq successfully

identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes. Int. J. Cancer.

Braun, T.A., Mullins, R.F., Wagner, A.H., Andorf, J.L., Johnston, R.M., Bakall, B.B., Deluca, A.P., Fishman, G.A., Lam, B.L., Weleber, R.G., *et al.* (2013). Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. Hum. Mol. Genet. *22*, 5136–5145.

Brillen, A.-L., Schöneweis, K., Walotka, L., Hartmann, L., Müller, L., Ptok, J., Kaisers, W., Poschmann, G., Stühler, K., Buratti, E., *et al.* (2017). Succession of splicing regulatory elements determines cryptic 5′ss functionality. Nucleic Acids Res. *45*, 4202–4216.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. Mol. Cell. Biol. *24*, 10505–10514.

Caputo, S., Benboudjema, L., Sinilnikova, O., Rouleau, E., Béroud, C., Lidereau, R., and French BRCA GGC Consortium (2012). Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. Nucleic Acids Res. *40*, D992-1002.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Castaman, G., Giacomelli, S.H., Mancuso, M.E., Sanna, S., Santagostino, E., and Rodeghiero, F. (2010). F8 mRNA studies in haemophilia A patients with different splice site mutations. Haemoph. Off. J. World Fed. Hemoph. *16*, 786–790.

Chasin, L.A. (2007). Searching for splicing motifs. Adv. Exp. Med. Biol. *623*, 85–106.

Cieply, B., and Carstens, R.P. (2015). Functional roles of alternative splicing factors in human disease. Wiley Interdiscip. Rev. RNA *6*, 311–326.

Colombo, M., Blok, M.J., Whiley, P., Santamariña, M., Gutiérrez-Enríquez, S., Romero, A., Garre, P., Becker, A., Smith, L.D., De Vecchi, G., *et al.* (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. Hum. Mol. Genet. *23*, 3666–3680.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., *et al.* (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*.

Davis, R.L., Homer, V.M., George, P.M., and Brennan, S.O. (2009). A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. Hum. Mutat. *30*, 221–227.

De Gasperi, R., Gama Sosa, M.A., Sartorato, E.L., Battistini, S., MacFarlane, H., Gusella, J.F., Krivit, W., and Kolodny, E.H. (1996). Molecular heterogeneity of late-onset forms of globoid-cell leukodystrophy. Am. J. Hum. Genet. *59*, 1233–1242.

Dear, A., Daly, J., Brennan, S.O., Tuckfield, A., and George, P.M. (2006). An intronic mutation within FGB (IVS1+2076 a-->g) is associated with afibrinogenemia and recurrent transient ischemic attacks. J. Thromb. Haemost. JTH *4*, 471–472.

Dhir, A., and Buratti, E. (2010). Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. FEBS J. *277*, 841–855.

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., and Martins, A. (2013a). 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. *34*, 1547–1557.

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., and Martins, A. (2013b). 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. *34*, 1547–1557.

D'Souza, I., and Schellenberg, G.D. (2000). Determinants of 4-repeat tau expression. Coordination between enhancing and inhibitory splicing sequences for exon 10 inclusion. J. Biol. Chem. *275*, 17700–17709.

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J.O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res. *42*, 10681–10697.

Faà, V., Incani, F., Meloni, A., Corda, D., Masala, M., Baffico, A.M., Seia, M., Cao, A., and Rosatelli, M.C. (2009). Characterization of a disease-associated mutation affecting a putative splicing regulatory element in intron 6b of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. J. Biol. Chem. *284*, 30024–30031.

Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature.

Fokkema, I.F.A.C., Taschner, P.E.M., Schaafsma, G.C.P., Celli, J., Laros, J.F.J., and den Dunnen, J.T. (2011). LOVD v.2.0: the next generation in gene variant databases. Hum. Mutat. *32*, 557–563.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., *et al.* (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.-P., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. U. S. A. *102*, 16176–16181.

Frebourg, T. (2014). The challenge for the next generation of medical geneticists. Hum. Mutat. *35*, 909–911.

Fu, X.-D. (2004). Towards a splicing code. Cell *119*, 736–738.

Gao, L., Wang, J., Wang, Y., and Andreadis, A. (2007). SR protein 9G8 modulates splicing of tau exon 10 via its proximal downstream intron, a clustering region for frontotemporal dementia mutations. Mol. Cell. Neurosci. *34*, 48–58.

Goedert, M., and Jakes, R. (1990). Expression of separate isoforms of human tau protein: correlation with the tau pattern in brain and effects on tubulin polymerization. EMBO J. *9*, 4225–4230.

Goedert, M., Spillantini, M.G., Jakes, R., Rutherford, D., and Crowther, R.A. (1989). Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. Neuron *3*, 519–526.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Grandval, P., Fabre, A.J., Gaildrat, P., Baert-Desurmont, S., Buisine, M.-P., Ferrari, A., Wang, Q., Béroud, C., and Olschwang, S. (2013). UMD-MLH1/MSH2/MSH6 databases: description and analysis of genetic variations in French Lynch syndrome families. Database J. Biol. Databases Curation *2013*, bat036.

Green, M.R. (1991). Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Annu. Rev. Cell Biol. *7*, 559–599.

Grodecká, L., Buratti, E., and Freiberger, T. (2017a). Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? Int. J. Mol. Sci. *18*.

Grodecká, L., Hujová, P., Kramárek, M., Kršjaková, T., Kováčová, T., Vondrášková, K., Ravčuková, B., Hrnčířová, K., Souček, P., and Freiberger, T. (2017b). Systematic analysis of splicing defects in selected primary immunodeficiencies-related genes. Clin. Immunol. Orlando Fla *180*, 33–44.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, *et al.* (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Haque, A., Buratti, E., and Baralle, F.E. (2010). Functional properties and evolutionary splicing constraints on a composite exonic regulatory element of splicing in CFTR exon 12. Nucleic Acids Res. *38*, 647–659.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., and Pease, L.R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene *77*, 51–59.

Homolova, K., Zavadakova, P., Doktor, T.K., Schroeder, L.D., Kozich, V., and Andresen, B.S. (2010). The deep intronic c.903+469T>C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. Hum. Mutat. *31*, 437–444.

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., *et al.* (2012). 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. *33*, 1228–1238.

de la Hoya, M., Soukarieh, O., López-Perolio, I., Vega, A., Walker, L.C., van Ierland, Y., Baralle, D., Santamariña, M., Lattimore, V., Wijnen, J., *et al.* (2016). 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. *25*, 2256–2268.

Iqbal, K., Liu, F., and Gong, C.-X. (2016). Tau and neurodegenerative disease: the story so far. Nat. Rev. Neurol. *12*, 15–27.

Ishii, S., Nakao, S., Minamikawa-Tachino, R., Desnick, R.J., and Fan, J.-Q. (2002). Alternative splicing in the alpha-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. Am. J. Hum. Genet. *70*, 994–1002.

Kashima, T., Rao, N., and Manley, J.L. (2007). An intronic element contributes to splicing repression in spinal muscular atrophy. Proc. Natl. Acad. Sci. U. S. A. *104*, 3426–3431.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Ke, S., Anquetil, V., Zamalloa, J.R., Maity, A., Yang, A., Arias, M.A., Kalachikov, S., Russo, J.J., Ju, J., and Chasin, L.A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. *28*, 11–24.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.

King, K., Flinter, F.A., Nihalani, V., and Green, P.M. (2002). Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. Hum. Genet. *111*, 548–554.

Kosik, K.S., Crandall, J.E., Mufson, E.J., and Neve, R.L. (1989). Tau in situ hybridization in normal and Alzheimer brain: localization in the somatodendritic compartment. Ann. Neurol. *26*, 352–361.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., *et al.* (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., *et al.* (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862-868.

Leman, R., Gaildrat, P., Gac, G.L., Ka, C., Fichou, Y., Audrezet, M.-P., Caux-Moncoutier, V., Caputo, S.M., Boutry-Kryza, N., Léone, M., *et al.* (2018). 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.

Lewandowska, M.A., Stuani, C., Parvizpur, A., Baralle, F.E., and Pagani, F. (2005). Functional studies on the ATM intronic splicing processing element. Nucleic Acids Res. *33*, 4007–4015.

Liu, F., and Gong, C.-X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol. Neurodegener. *3*, 8.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. *579*, 1900–1903.

Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., Santamariña, M., Behar, R., Blanco, A., Carrasco, E., *et al.* (2018). Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? Front. Genet. *9*, 366.

Motohashi, K. (2015). A simple and efficient seamless DNA cloning method using SLiCE from Escherichia coli laboratory strains and its application to SLiP site-directed mutagenesis. BMC Biotechnol. *15*, 47.

Mueller, W.F., Larsen, L.S.Z., Garibaldi, A., Hatfield, G.W., and Hertel, K.J. (2015). The Silent Sway of Splicing by Synonymous Substitutions. J. Biol. Chem. *290*, 27700–27711.

Naftelberg, S., Schor, I.E., Ast, G., and Kornblihtt, A.R. (2015). Regulation of alternative splicing through coupling with transcription and chromatin structure. Annu. Rev. Biochem. *84*, 165–198.

Nathan, N., Girodon, E., Clement, A., and Corvol, H. (2012). A rare CFTR intronic mutation related to a mild CF disease in a 12-year-old girl. BMJ Case Rep. *2012*.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., and Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. *15*, 256–278.

Pagani, F., Buratti, E., Stuani, C., Bendix, R., Dörk, T., and Baralle, F.E. (2002). A new type of mutation causes a splicing defect in ATM. Nat. Genet. *30*, 426–429.

Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., and Baralle, F.E. (2003). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. Hum. Mol. Genet. *12*, 1111–1120.

Pagani, F., Raponi, M., and Baralle, F.E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. U. S. A. *102*, 6368–6372.

Palhais, B., Dembic, M., Sabaratnam, R., Nielsen, K.S., Doktor, T.K., Bruun, G.H., and Andresen, B.S. (2016). The prevalent deep intronic c. 639+919 G>A GLA mutation causes pseudoexon activation and Fabry disease by abolishing the binding of hnRNPA1 and hnRNP A2/B1 to a splicing silencer. Mol. Genet. Metab. *119*, 258–269.

Park, J.W., Jung, S., Rouchka, E.C., Tseng, Y.-T., and Xing, Y. (2016). rMAPS: RNA map analysis and plotting server for alternative exon regulation. Nucleic Acids Res. *44*, W333-338.

Pastor, T., and Pagani, F. (2011). Interaction of hnRNPA1/A2 and DAZAP1 with an Alu-derived intronic splicing enhancer regulates ATM aberrant splicing. PloS One *6*, e23349.

Pastor, T., Talotti, G., Lewandowska, M.A., and Pagani, F. (2009). An Alu-derived intronic splicing enhancer facilitates intronic processing and modulates aberrant splicing in ATM. Nucleic Acids Res. *37*, 7258–7267.

Plazzer, J.P., Sijmons, R.H., Woods, M.O., Peltomäki, P., Thompson, B., Den Dunnen, J.T., and Macrae, F. (2013). The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. Fam. Cancer *12*, 175–180.

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. J. Hum. Genet. *59*, 5–15.

Raponi, M., Kralovicova, J., Copson, E., Divina, P., Eccles, D., Johnson, P., Baralle, D., and Vorechovsky, I. (2011). Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. Hum. Mutat. *32*, 436–444.

Rhine, C.L., Cygan, K.J., Soemedi, R., Maguire, S., Murray, M.F., Monaghan, S.F., and Fairbrother, W.G. (2018). Hereditary cancer genes are highly susceptible to splicing mutations. PLoS Genet. *14*, e1007231.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., *et al.* (2015). 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. *17*, 405–424.

Rincón, A., Aguado, C., Desviat, L.R., Sánchez-Alcudia, R., Ugarte, M., and Pérez, B. (2007). Propionic and methylmalonic acidemia: antisense therapeutics for intronic variations causing aberrantly spliced messenger RNA. Am. J. Hum. Genet. *81*, 1262–1270.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Sabbagh, A., Pasmant, E., Imbard, A., Luscan, A., Soares, M., Blanché, H., Laurendeau, I., Ferkal, S., Vidaud, M., Pinson, S., *et al.* (2013). NF1 molecular characterization and neurofibromatosis type I genotype-phenotype correlation: the French experience. Hum. Mutat. *34*, 1510–1518.

Saha, A., Kim, Y., Gewirtz, A.D.H., Jo, B., Gao, C., McDowell, I.C., GTEx Consortium, Engelhardt, B.E., and Battle, A. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. *27*, 1843–1858.

Sanz, D.J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardeñosa, E., Lastra, E., Pagani, F., Miner, C., and Velasco, E.A. (2010). A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *16*, 1957–1967.

Savisaar, R., and Hurst, L.D. (2017). Estimating the prevalence of functional exonic splice regulatory information. Hum. Genet. *136*, 1059–1078.

Savisaar, R., and Hurst, L.D. (2018). Exonic splice regulation imposes strong selection at synonymous sites. Genome Res. *28*, 1442–1454.

Schulz, H.L., Grassmann, F., Kellner, U., Spital, G., Rüther, K., Jägle, H., Hufendiek, K., Rating, P., Huchzermeyer, C., Baier, M.J., *et al.* (2017). Mutation Spectrum of the ABCA4 Gene in 335 Stargardt Disease Patients From a Multicenter German Cohort-Impact of Selected Deep Intronic Variants and Common SNPs. Invest. Ophthalmol. Vis. Sci. *58*, 394–403.

Shendure, J. (2011). Next-generation human genetics. Genome Biol. *12*, 408.

Shepard, P.J., Choi, E.-A., Busch, A., and Hertel, K.J. (2011). Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites. Nucleic Acids Res. *39*, 8928–8937.

Singh, N.N., Singh, R.N., and Androphy, E.J. (2007). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Res. *35*, 371–389.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., Tosi, M., and Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using *In silico* Tools. PLoS Genet. *12*, e1005756.

Spena, S., Asselta, R., Platé, M., Castaman, G., Duga, S., and Tenchini, M.L. (2007). Pseudo-exon activation caused by a deep-intronic mutation in the fibrinogen gamma-chain gene as a novel mechanism for congenital afibrinogenaemia. Br. J. Haematol. *139*, 128–132.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K., and Cooper, D.N. (2012). The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr. Protoc. Bioinforma. *Chapter 1*, Unit1.13.

Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum. Genet. *136*, 665–677.

Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N., and Sanford, J.R. (2011). Loss of exon identity is a common mechanism of human inherited disease. Genome Res. *21*, 1563–1571.

Strobel, E.J., Yu, A.M., and Lucks, J.B. (2018). High-throughput determination of RNA structures. Nat. Rev. Genet. *19*, 615–634.

Sun, L., Fazal, F.M., Li, P., Broughton, J.P., Lee, B., Tang, L., Huang, W., Kool, E.T., Chang, H.Y., and Zhang, Q.C. (2019). RNA structure maps across mammalian cellular compartments. Nat. Struct. Mol. Biol. *26*, 322–330.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.

Szabo, C., Masiello, A., Ryan, J.F., and Brody, L.C. (2000). The breast cancer information core: database design, structure, and scope. Hum. Mutat. *16*, 123–131.

Tammaro, C., Raponi, M., Wilson, D.I., and Baralle, D. (2014). BRCA1 EXON 11, a CERES (composite regulatory element of splicing) element involved in splice regulation. Int. J. Mol. Sci. *15*, 13045–13059.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J., *et al.* (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. *29*, 1412–1424.

Trabelsi, M., Beugnet, C., Deburgrave, N., Commere, V., Orhant, L., Leturcq, F., and Chelly, J. (2014). When a mid-intronic variation of DMD gene creates an ESE site. Neuromuscul. Disord. NMD *24*, 1111–1117.

Vandeweyer, G., Van Laer, L., Loeys, B., Van den Bulcke, T., and Kooy, R.F. (2014). VariantDB: a flexible annotation and filtering portal for next generation sequencing data. Genome Med. *6*, 74.

Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. Hum. Genet. *136*, 1093–1111.

Wainberg, M., Alipanahi, B., and Frey, B. (2016). Does conservation account for splicing patterns? BMC Genomics *17*, 787.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA N. Y. N *14*, 802–813.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2015a). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2015b). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yang, E.-W., Bahn, J.H., Hsiao, E.Y.-H., Tan, B.X., Sun, Y., Fu, T., Zhou, B., Van Nostrand, E.L., Pratt, G.A., Freese, P., *et al.* (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. Nat. Commun. *10*, 1338.

Yee, B.A., Pratt, G.A., Graveley, B.R., Van Nostrand, E.L., and Yeo, G.W. (2019). RBP-Maps enables robust generation of splicing regulatory maps. RNA N. Y. N *25*, 193–204.

Zavadáková, P., Fowler, B., Suormala, T., Novotna, Z., Mueller, P., Hennermann, J.B., Zeman, J., Vilaseca, M.A., Vilarinho, L., Gutsche, S., *et al.* (2005). cblE type of homocystinuria due to methionine synthase reductase deficiency: functional correction by minigene expression. Hum. Mutat. *25*, 239–247.

Legends to figures and tables

Figure 1. Small-scale assessment of the performance of SRE-dedicated bioinformatics predictions by using two previously described datasets. (A) Mutational landscape of the MMR dataset used by Xiong *et al.*, 2015 to evaluate SPANR-based *in silico* predictions (n=225 *MLH1* and *MSH2* variants and their corresponding RNA splicing data retrieved from Xiong *et al.*, 2015). The panel shows the distribution of *MMR* variants depending on their position relative to the nearest 5' or 3' splice site and their impact on splicing as determined experimentally. (B) Receiver operating characteristic (ROC) curve analysis of the SPANR approach in predicting exon skipping events within the MMR dataset. Areas under the curve (AUC) are provided at 99% CI. (C) Comparative statistical analysis of the predictive power of four SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and the newly developed HAL) by using the pilot dataset described in Soukarieh *et al.*, 2016 (n= 154 variants distributed within 5 exons and their corresponding RNA data). Statistical analyses were performed as described under Materials and Methods. Results are expressed as two-sided p-values (*** p-value<0.001). The number of variants (n=) taken into consideration in each statistical analysis is highlighted in grey. Variants for which

semi-quantitative RNA information was not available were excluded from severity correlation analyses. n/s, not statistically significant. (D) ROC curves of the four SRE-dedicated approaches in predicting exon-skipping events within the pilot dataset described in Soukarieh *et al.*, 2016. Areas under the curve (AUC) are provided at 99% CI. (E) Characterization of the performance of SRE-dedicated tools in predicting variant-induced exon skipping within the pilot dataset. True and false calls were determined by taking into account the thresholds indicated between parentheses under each method: -0.50 for QUEPASA, -20 for HEXplorer, and -5% for SPANR (as reported in Soukarieh *et al.* 2016), or -0.38% for SPANR and -5% for HAL (as inferred from the SPANR ROC curve in (D), and recommended by Rosenberg *et al.*, 2015, respectively).

Figure 2. Large-scale evaluation of the performance of SRE-dedicated *in silico* tools in predicting variant-induced exon skipping by using an extensive training dataset. (A) Schematic representation of variant-induced exon skipping events. (B) Statistical analysis of the performance of the four SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and HAL) in predicting exon skipping by using a training dataset containing splicing data on 1214 variants distributed within 190 exons from 88 different genes (Supplementary Table S3). Statistical analyses were performed as described under Materials and Methods. Results are expressed as two-sided p-values (*** p-value<0.001). The number of variants (n=) taken into consideration in each statistical analysis is highlighted in grey. Variants for which semi-quantitative RNA information was not available were excluded from severity correlation analyses. (C) Receiver operating characteristic (ROC) curve analysis of the four SRE-dedicated approaches in predicting variant-induced exon skipping within the training set. Areas under the curve (AUC) are provided at 99% CI. (D) Performance characterization of the SRE-dedicated tools in predicting variant-induced exon skipping. True and false calls were determined by taking into account optimal thresholds inferred from the ROC curves shown in (C). The newly established thresholds for predicting exon skipping are indicated between parentheses under each method. (E) Venn diagram illustrating the number of false calls (FC) produced by each SRE-dedicated approach (and their overlap) in predicting variant-induced exon skipping within the training dataset. (F) Jackknife evaluation of the performance of each stand-alone SRE-dedicated method and all their possible combinations in predicting variant-induced exon skipping within the training set. The arrows point to the combinations showing the best compromises between sensitivity and specificity (QUEPASA&HAL, at least 3, and LR_{skip}). (G) Venn diagram illustrating the number of false calls

(FC) produced by the three best combinations of SRE-dedicated approaches) in predicting variant-induced exon skipping as pointed in (F). A more detailed comparative performance assessment is shown in (D).

Figure 3. Large-scale evaluation of the performance of SRE-dedicated *in silico* tools in predicting variant-increased exon inclusion by using an extensive training dataset. (A) Schematic representation of variant-increased exon inclusion events. (B) Statistical analysis of the performance of the four SRE-dedicated bioinformatics approaches (QUEPASA, HEXplorer, SPANR and HAL) in predicting increased exon inclusion by using a training dataset containing 1214 variants distributed within 190 exons from 88 different genes (Supplementary Table S4). Statistical analyses were performed as described under Materials and Methods. Results are expressed as two-sided p-values (*** p-value<0.001). The number of variants (n=) taken into consideration in each statistical analysis is highlighted in grey. (C) Receiver operating characteristic (ROC) curve analysis of SRE-dedicated approaches in predicting variant-increased exon inclusion within the training set. Areas under the curve (AUC) are provided at 99% CI. (D) Performance characterization of the SRE-dedicated bioinformatics tools in predicting variant-increased exon inclusion within the training set. True and false calls were determined by taking into account optimal thresholds inferred from the ROC curves shown in (C). The newly established thresholds for predicting exon inclusion are indicated between parentheses under each method. (E) Venn diagram illustrating the number of false calls (FC) produced by each SRE-dedicated approach (and their overlap) in predicting variant-increased exon inclusion. (F) Performance of each stand-alone SRE-dedicated method and all their possible combinations in predicting variant-increased exon inclusion obtained from Jackknife cross-validation on the training set. The arrows point to the combinations showing the best compromises between sensitivity and specificity (QUEPASA&HAL, at least 3, and LR_{inc}). (G) Venn diagram illustrating the number of false calls (FC) produced by the three combinations highlighted in (F). A more detailed comparative performance assessment is presented in (D).

Figure 4. A prospective study identifies new disease-associated splicing mutations and validates SRE-dedicated *in silico* approaches for predicting variant-induced splicing alterations. (A) Mutational landscape of the prospective validation dataset containing 150 variants identified in three exons from disease-associated genes, all mapping to positions outside the

reference 3' and 5'ss. The following colour code reflects the impact on splicing of each variant relative to wild-type: red, increased exon skipping; black, no effect on splicing; green, increased exon inclusion. These effects, which are further detailed in Supplementary Table S5, were experimentally determined by performing cell-based minigene splicing assays as described under Materials and Methods. (B) Evaluation of the performance of the SRE-dedicated *in silico* approaches in predicting variant-induced exon skipping within the validation dataset. The predictive scores produced by each *in silico* method are shown in Supplementary Table S5. Decision thresholds for predictions of exon skipping were derived from the analysis of the training dataset and are shown between parentheses for each SRE-dedicated approach (-0.5 for QUEPASA, -14 for HEXplorer, -0.1% for SPANR, -3.4% for HAL, and +31.1% for LR_{skip}). (C) Evaluation of the performance of the SRE-dedicated *in silico* approaches in predicting variant-induced exon inclusion within the validation dataset. The predictive scores produced by each *in silico* method are shown in Supplementary Table S5. Decision thresholds for predictions of exon inclusion were derived from the analysis of the training dataset and are shown between parentheses for each SRE-dedicated method (+0.36 for QUEPASA, +9 for HEXplorer, +0.3% for SPANR, +1.0% for HAL, and +6.2% for LR_{inc}.).

Figure S1. Datasets used in this study for evaluating the predictive power of SRE-dedicated *in silico* approaches(A) Outline of the five data collections used in this work including two previously reported datasets (MMR dataset and pilot dataset) and three newly established datasets (training dataset, validation dataset, and preliminary pEX-SRE dataset) all described under Materials and Methods. The RNA data from the validation dataset were generated in this study (prospective work) whereas those of the other datasets were retrieved from the literature (retrospective work) (B) Overview of the training dataset. This set contains 1214 exonic variants distributed within 190 exons from 88 different genes (all located outside reference 3' and 5'ss), which were divided into 22 subsets depending on which exon/gene they map to, including 15 exon-specific subsets, 6 gene-specific subsets, and 1 subset here named "others" (please see Supplementary Table S4 for details). Exon-specific subsets were created when there was RNA data for at least 10 variants per exon, some of which having an impact on splicing. Variants in exons with less than 10 variants/exon, and having different consequences on splicing were pooled together into gene-specific subsets (*ABCB11*, *BRCA1*, *BRCA2*, *DYSF*, *MLH1* and *MSH2*), whereas the remaining variants were included in the "others" subset. As indicated in the right column, in

addition to qualitative information on RNA splicing (type of splicing effects), 9 subsets also contained semi-quantitative data for most of the corresponding variants (described in Supplementary Table S4).

Figure S2. Selection of 22 *MSH2* exon 5 variants for a prospective study and integration into the validation dataset. First, we retrieved from human variation databases all nucleotide variants reported in *MSH2* exon 5 (n=112) (Supplementary Table S6). Only exonic substitutions located outside reference splice sites (n=75) were retained for an exploratory SRE-dedicated *in silico* analysis with QUEPASA, HEXplorer, SPANR and HAL (Supplementary Table S6). We then applied the "at least 3" decision rule with the thresholds inferred from the training dataset to predict which variants were more likely to induce exon skipping and those less likely to do so (panels A and B, respectively). Next, an equivalent number of variants were selected from these 2 groups (n=11 each) by taking into account the 5 SNVs with the most extreme negative or positive scores produced by each SRE-dedicated approach as indicated in the tables at the bottom. The grey background indicates variants present in a precedent column in the same table. Variants in the white background thus represent all those selected for experimental analyses and integration into the validation dataset.

Figure S3. Structure of the pCAS2 and pSPL3mK constructs used in the cell-based minigene splicing assays. (A) pCAS2 minigene vector. The pCAS2 vector carries two exons (A1 and B1) with a sequence derived from the human *SERPING1/C1NH* gene, separated by an intron containing BamHI and MluI cloning sites. As described in Soukarieh *et al.*, 2016, the exon A1 of pCAS2 contains a disrupted version of the translation initiation codon of *SERPING1/C1NH*. Expression of the pCAS2 minigene is under the control of a CMV promoter. (B) pSPL3mK minigene vector. The pSPL3mK plasmid carries two chimeric exons (here named A2 and B2, both containing rabbit β-globin (βg) and HIV Tat sequences, as indicated) separated by an intron containing BamHI and MluI cloning sites. This vector is a modified version of the pSPL3m plasmid described by Tournier *et al.*, 2008. Briefly, here we prepared the pSPL3mK construct by disrupting the rabbit b-globin translation initiation codon present in pSPL3m exon I by replacing the ATG sequence with GTG. Expression of the pSPL3mK minigene is driven by the SV40 promoter. PA, polyadenylation site**.**

Figure S4. Comparison of the variant-associated splicing effects described in the training dataset with *in silico* data obtained with SRE-dedicated approaches. The 1214 variations from

the training dataset were separated into 3 groups according to their impact on splicing as experimentally determined and described in Supplementary Table S4. Panels A to F compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL, LR_{skip} and LR_{inc}, respectively (Supplementary Table S4). Two-sided p-values were calculated by using ANOVA or Kruskal-Wallis, as indicated in Supplementary Table S10.

Figure S5. Correlation between variant-associated exon inclusion levels described in the training dataset and *in silico* data obtained with SRE-dedicated approaches. Exon inclusion levels refer to semi-quantitative data available from minigene splicing assays for 9 out of the 22 subsets of variants described in the training dataset (Supplementary Figure S1 and Supplementary Table S4). Panels A to F compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL, LR_{skip} and LR_{inc}, respectively Supplementary Table S3). Determination coefficients (R^2) and two-sided p-values were determined by performing a Pearson or Spearman correlation analysis, as indicated in Supplementary Table S10.

Figure S6. Comparison of the variant-associated splicing effects observed in the validation dataset with *in silico* data obtained with SRE-dedicated approaches. The 150 variations from the validation dataset were separated into 3 groups according to their impact on splicing as experimentally determined by performing minigene splicing assays (Figure 4 and Supplementary Table S5). Panels A to F compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL, LR_{skip} and LR_{inc}, respectively (Supplementary Table S5). Two-sided p-values were calculated by using ANOVA or Kruskal-Wallis, as indicated in Supplementary Table S10.

Figure S7. Correlation between variant-associated exon inclusion levels observed in the validation dataset and *in silico* data obtained with SRE-dedicated approaches. Exon inclusion levels refer to semi-quantitative results obtained in minigene splicing assays performed for the 150 variants of the validation dataset distributed in 3 different exons as indicated (Figure 4 and Supplementary Table S5). Panels A to F compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL, LR_{skip} and LR_{inc}, respectively (Supplementary Table S5). Determination coefficients (R^2) and two-sided p-values were determined by performing a Pearson or Spearman correlation analysis, as indicated in Supplementary Table S10.

Figure S8. Prediction of pseudoexon inclusion triggered by deep intronic variants suspected to alter splicing regulatory elements. (A) Representative example of a deep intronic variant (c.385-719G>A) that leads to the inclusion of a 147 nt pseudoexon in *COL4A5* mRNA by creating a binding site for the splicing activator SRSF1 protein (ESE) (King *et al.*, 2002). Grey boxes represent natural exons whereas the white box indicates a variant-induced pseudoexon created deep in the intron via the activation of flanking cryptic 3' and 5'ss. Numbers below the natural and the cryptic AG/GT splice sites refer to MES scores (identical in the WT and variant contexts). (B) Representative example of a deep intronic variant (c.640-801G>A) leading to inclusion of a 57 nt pseudoexon in GLA mRNA by probably disrupting a splicing silencer that binds the splicing repressor hnRNPA1/A2 protein (ESS) (Palhais *et al.*, 2016). (C) Preliminary pEX-SRE dataset and comparison with SRE-dedicated bioinformatics predictions. The pEX-SRE dataset contains 13 disease-causing deep intronic variants known to create pseudoexons (pEX), as determined experimentally, by altering potential SREs. This collection was compiled from the literature as described under Materials and Methods. The table shows the identity of each variant as well as their distance relative to pseudoexon boundaries and the corresponding literature references. *In silico* predictions of potential variant-induced splicing alterations were conducted by using QUEPASA and HEXplorer, the only new SRE-dedicated approaches that can be applied to deep intronic variants (Supplementary Table S1). True and false calls of pseudoexon creation (highlighted in white and grey, respectively) were determined by taking into account the optimal thresholds inferred from the analysis of the training dataset for predicting variant-induced exon inclusion : +0.36 for QUEPASA and +9 for HEXplorer. [1] Braun *et al.*, 2013, [2] Bauwens *et al.*, 2015, [3] Schulz *et al.*, 2017, [4] Pagani *et al.*, 2002, [5] Lewandowska *et al.*, 2005, [6] Dhir and Buratti, 2010, [7] Pastor *et al.*, 2009, [8] Pastor and Pagani, 2011, [9] Faà *et al.*, 2009, [10] Nathan *et al.*, 2012, [11] King *et al.*, 2002, [12] Trabelsi *et al.*, 2014, [13] Castaman *et al.*, 2010, [14] Davis *et al.*, 2009, [15] Spena *et al.*, 2007, [16] Dear *et al.*, 2006, [17] De Gasperi *et al.*, 1996, [18] (shii *et al.*, 2002, [19] Palhais *et al.*, 2016, [20] Homolova *et al.*, 2010, [21] Zavadáková *et al.*, 2005, [22] Sabbagh *et al.*, 2013, [23] Rincón *et al.*, 2007.

Figure S9. Large-scale evaluation of the performance of SRE-dedicated *in silico* tools in predicting variant-induced splicing alterations at 95% specificity or sensitivity by using an extensive training dataset. (A) Jackknife evaluation of the performance of each stand-alone SRE-dedicated method and all their possible combinations in predicting variant-induced exon skipping

within the training set by taking into account thresholds that yields to 95% sensitivity. (B) Jackknife evaluation of the performance of each stand-alone SRE-dedicated method and all their possible combinations in predicting variant-increased exon inclusion within the training set by taking into account thresholds that yields to 95% sensitivity. (C) Performance characterization of the SRE-dedicated tools in predicting variant-induced exon skipping and variant-increased exon inclusion by taking into account thresholds that yields to 95% specificity. True and false calls were determined by taking into account thresholds inferred from the ROC curves shown in Figures 2C and 3C that yields to 95% sensitivity. The newly established thresholds for predicting exon skipping or exon inclusion are indicated under each method. (D) Jackknife evaluation of the performance of each stand-alone SRE-dedicated method and all their possible combinations in predicting variant-induced exon skipping within the training set by taking into account thresholds that yields to 95% specificity. (E) Jackknife evaluation of the performance of each stand-alone SRE-dedicated method and all their possible combinations in predicting variant-increased exon inclusion within the training set by taking into account thresholds that yields to 95% specificity. (F) Performance characterization of the SRE-dedicated tools in predicting variant-induced exon skipping and variant-increased exon inclusion by taking into account thresholds that yields to 95% specificity. True and false calls were determined by taking into account thresholds inferred from the ROC curves shown in Figures 2C and 3C that yields to 95% specificity. The newly established thresholds for predicting exon skipping or exon inclusion are indicated under each method.

Table S1. Overview of the four SRE-dedicated *in silico* approaches evaluated in this study. [1] Ke *et al.*, 2011; [2] Di Giacomo *et al.*, 2013; [3] Erkelenz *et al.*, 2014; [4] Xiong *et al.*, 2015; [5] Rosenberg *et al.*, 2015. Inc, Inclusion ; PSI, percent spliced in; Skip, Skipping; VCF, variant called format; WT, wild-type

Table S2. MMR dataset and associated SRE-dedicated *in silico* predictions This dataset contains RNA splicing data on 225 variants identified in *MLH1* and *MSH2* (MMR genes) and was recently used by Xiong *et al.* to evaluate the performance of SPANR in predicting variant-induced splicing defects (Xiong *et al.*, 2015). The table indicates the impact on splicing of each variant, as well as their position relative to the nearest splice site and the scores obtained by performing *in silico* analyses with MaxEntScan (MES and ΔMES scores), SpliceSiteFinder-Like (SSFL and ΔSSFL scores) and SPANR (ΔΨ scores, %), as described under Materials and Methods.

Table S3. Pilot dataset and associated SRE-dedicated *in silico* predictions. This dataset contains RNA splicing data on 154 variants, which are distributed within 5 exons from 5 different genes, as indicated, and was recently used by Soukarieh *et al.* to evaluate the performance of QUEPASA, HEXplorer and SPANR in predicting splicing defects caused by exonic variants located outside splice sites (Soukarieh *et al.* 2016). The table indicates the impact on splicing of each variant as determined experimentally, as well as the scores obtained by performing *in silico* analyses with QUEPASA (ΔtERSseq scores), HEXplorer (ΔHZ_{EI} scores), SPANR (ΔΨ scores, %) and HAL (ΔΨ scores, %), as described under Materials and Methods.

Table S4. Training dataset and associated SRE-dedicated *in silico* predictions. The training dataset contains RNA splicing data on 1214 variants distributed within 190 exons from 88 disease-associated genes. This collection represents a largely extended version of the pilot dataset and was prepared by querying the literature for exonic variants for which RNA data were available and then excluding those directly mapping to splice sites, as described under Materials and Methods. The table indicates the impact on splicing of each variant, the sources of RT-PCR data (minigene splicing assays or patients' RNA analysis) and the corresponding bibliographic references, as well as the scores obtained in this study by performing *in silico* analyses with QUEPASA (ΔtERSseq scores), HEXplorer (ΔHZ_{EI} scores), SPANR (ΔΨ scores, %), HAL (ΔΨ scores, %), LR_{skip} (%) and LR_{Inc} (%). When information on the level of WT exon inclusion was not available, the value entered into the HAL interface was set to 99%. n/a, not applicable; nd, not determined.

Table S5. Validation dataset and associated SRE-dedicated *in silico* predictions. The validation dataset contains RNA data generated in this study for 150 variants distributed within 3 exons from 3 disease-associated genes (*MSH2* exon 5, *BRCA1* exon 5 and *MAPT* exon 10). These variants were retrieved from human variation databases and none map to splice sites. Their impact on RNA splicing was assessed by performing minigene splicing assays, either with pCAS2- or pSPL3mK-derived minigenes, as described under Materials and Methods. The table describes the experimental results produced in this study, as well as the scores obtained by performing *in silico* analyses with QUEPASA (ΔtERSseq scores), HEXplorer (ΔHZ_{EI} scores), SPANR (ΔΨ scores, %), HAL (ΔΨ scores, %), LR_{skip} (%) and LR_{Inc} (%). The bibliographic references indicate variants for which independent RNA data is available in the literature. n/a, not applicable

Table S6. Extraction of nucleotide variants mapping to *MSH2* exon 5 reported in human variation databases and integration of 22 variants into the validation dataset. First, all *MSH2* exon 5 nucleotide changes were retrieved from human variation databases (n=112 variants). After eliminating variants mapping to the reference 3' and 5'ss (indicated by the black background), we then performed an exploratory SRE-dedicated *in silico* analysis by using QUEPASA (ΔtERSseq scores), HEXplorer (ΔHZ_{EI} scores), SPANR (ΔΨ scores, %), and HAL (ΔΨ scores, %), as indicated. Finally, 22 single nucleotide substitutions were selected for integrating the validation dataset by including 11 variants predicted as the most susceptible to induce exon skipping (indicated in red) and the other 11 as the less likely to do so (indicated in green) according to the "at least 3" decision rule and further described under Supplementary Figure S2. The bibliographic references indicate variants for which independent RNA data is available in the literature.

Table S7. Extraction of nucleotide variants mapping to *BRCA1* exon 5 reported in human variation databases and integration of 98 variants into the validation dataset. First, all *BRCA1* exon 5 nucleotide changes were retrieved from human variation databases (n=106 variants) and their impact on splicing was evaluated by performing minigene splicing assays, as described under Materials and Methods. Variants mapping to the reference 3' and 5'ss or impairing the alternative 5'ss of *BRCA1* exon 5 (indicated by the black and the grey backgrounds, respectively) were excluded from further analyses. The remaining 98 variations were selected for integrating the validation dataset. SRE-dedicated *in silico* analyses were performed for each variant by using QUEPASA (ΔtERSseq scores), HEXplorer (ΔHZ_{EI} scores), SPANR (ΔΨ scores, %), and HAL (ΔΨ scores, %), as shown. The bibliographic references indicate variants for which independent RNA data is available in the literature.

Table S8. Extraction of nucleotide variants mapping to *MAPT* exon 10 reported in human variation databases and integration of 30 variants into the validation dataset. First, all *MAPT* exon 10 nucleotide changes were retrieved from human variation databases (n=34 variants). Then, variants mapping to the reference 3' and 5'ss (indicated by the black background) were excluded from further analyses, whereas the remaining 30 variants were retained for integrating the validation dataset. SRE-dedicated *in silico* analyses were performed by using QUEPASA (ΔtERSseq scores), HEXplorer (ΔHZ_{EI} scores), SPANR (ΔΨ scores, %), and HAL (ΔΨ scores, %),

as described under Materials and Methods. The bibliographic references indicate variants for which independent RNA data is available in the literature.

Table S9. Primers used in the minigene splicing reporter assays.

1 F, forward; R, reverse.

2 Intronic and exonic sequences are indicated in grey and black, respectively. The position of the nucleotide variant is underlined. The double underlined sequences correspond to restriction sites for BamHI and MluI and the sequence highlighted in grey correspond to the 15bp-tail used for homologous recombination cloning.

Table S10. Statistical analyses conducted in this study. Data derived from confrontation of experimental and *in silico* analyses were compared by using either Student's test, one-way ANOVA test and Pearson's correlation or their derivatives depending on the purpose of the analysis and data distribution patterns, as indicated.

Table S11. Comparative analysis of the performance of SRE-dedicated bioinformatics approaches in predicting variant-induced exon skipping in each of subset of the training and validation datasets. The experimental data of each subset of the training and validation datasets (indicated in the left column) were compared with the corresponding *in silico* results obtained with the four stand-alone SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and HAL) and their combinations (QUEPASA&HAL, at least 3, and LR_{Skip}) (Supplementary Tables S4 and S5). The asterisks indicate which subsets belong to the validation dataset. True and false calls of variant-induced exon skipping were determined by taking into account thresholds inferred from the ROC curve analysis of the full training dataset: -0.50 for QUEPASA, -14 for HEXplorer, -0.1% for SPANR, -3.4% for HAL and 31.1% for LR_{skip}. Sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV) and accuracy (Acc) were calculated as described under Materials and Methods. n=, number of variants taken into consideration in each analysis (those referring to SPANR and LR_{skip} are indicated between parentheses).

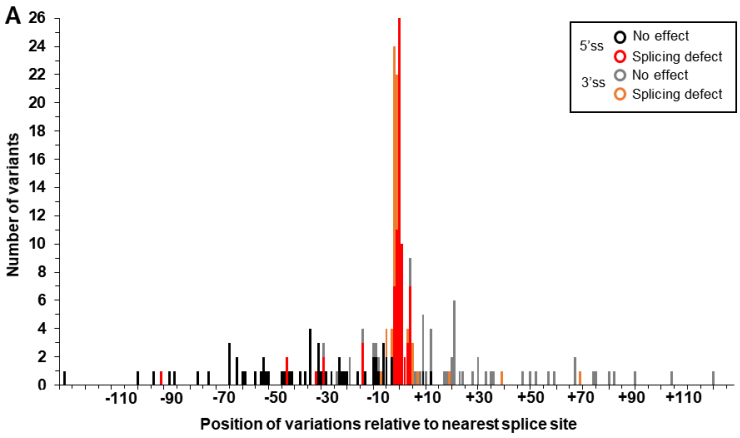
Table S12. Comparative analysis of the performance of SRE-dedicated bioinformatics approaches in predicting variant-increased exon inclusion in subsets of the training and validation datasets. The experimental data of subsets of the training and validation datasets for

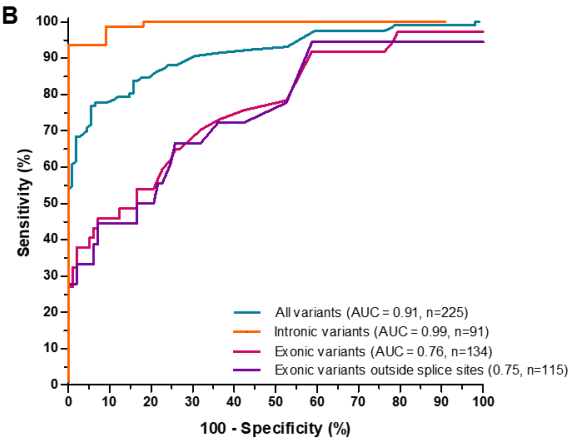
which there were cases of increased exon inclusion (indicated in the left column) were compared with the corresponding *in silico* results obtained with the four stand-alone SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and HAL) and their combinations (QUEPASA&HAL, at least 3, and LR_{Skip}) (Supplementary Table S4 and S5). The asterisks indicate which subsets belong to the validation dataset. True and false calls of variant-increased exon inclusion were determined by taking into account thresholds inferred from the ROC curve analysis of the full training dataset: +0.36 for QUEPASA, +9 for HEXplorer, +0.3% for SPANR, -1.0% for HAL and 6.2% for LR_{inc}. Sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV) and accuracy (Acc) were calculated as described under Materials and Methods. n=, number of variants taken into consideration in each analysis (those referring to SPANR and LR_{skip} are indicated between parentheses).

Table S13. Optimisation of the thresholds of SRE-dedicated approaches for predicting variant-induced exon skipping in each subset of the training and validation datasets. The experimental data of each subset of the training and validation datasets (indicated in the left column) were compared with the corresponding *in silico* results obtained with the four stand-alone SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and HAL) and their combinations (QUEPASA&HAL, at least 3, and LR_{Skip}) (Supplementary Table S4 and S5). The asterisks indicate which subsets belong to the validation dataset. True and false calls of variant-induced exon skipping were determined by taking into account the best threshold (Thr) inferred from individual ROC curves performed for each data subset and each algorithm. Sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV) and accuracy (Acc) were calculated as described under Materials and Methods. n=, number of variants taken into consideration in each analysis (those referring to SPANR and LR_{skip} are indicated between parentheses).

Table S14. Optimisation of the thresholds of SRE-dedicated bioinformatics approaches in predicting variant-increased exon inclusion in subsets of the training and validation datasets. The experimental data of subsets of the training and validation datasets for which there were cases of increased exon inclusion (indicated in the left column) were compared with the corresponding *in silico* results obtained with the four stand-alone SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and HAL) and their combinations (QUEPASA&HAL, at least 3, and LR_{Inc})

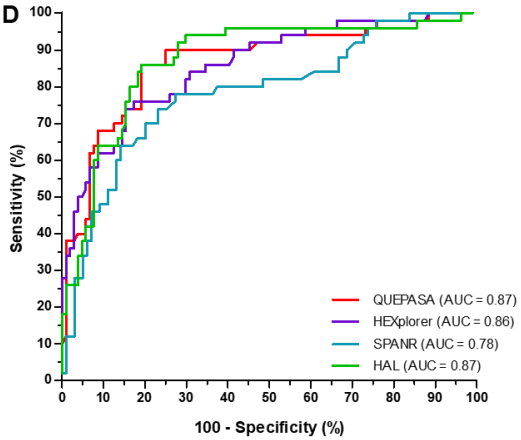
(Supplementary Tables S4 and S5). The asterisks indicate which subsets belong to the validation dataset. True and false calls of variant-increased exon inclusion were determined by taking into account the best threshold (Thr) inferred from individual ROC curves performed for each data subset and each algorithm. Sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV) and accuracy (Acc) were calculated as described under Materials and Methods. n=, number of variants taken into consideration in each analysis (those referring to SPANR and LR_{skip} are indicated between parentheses).





C

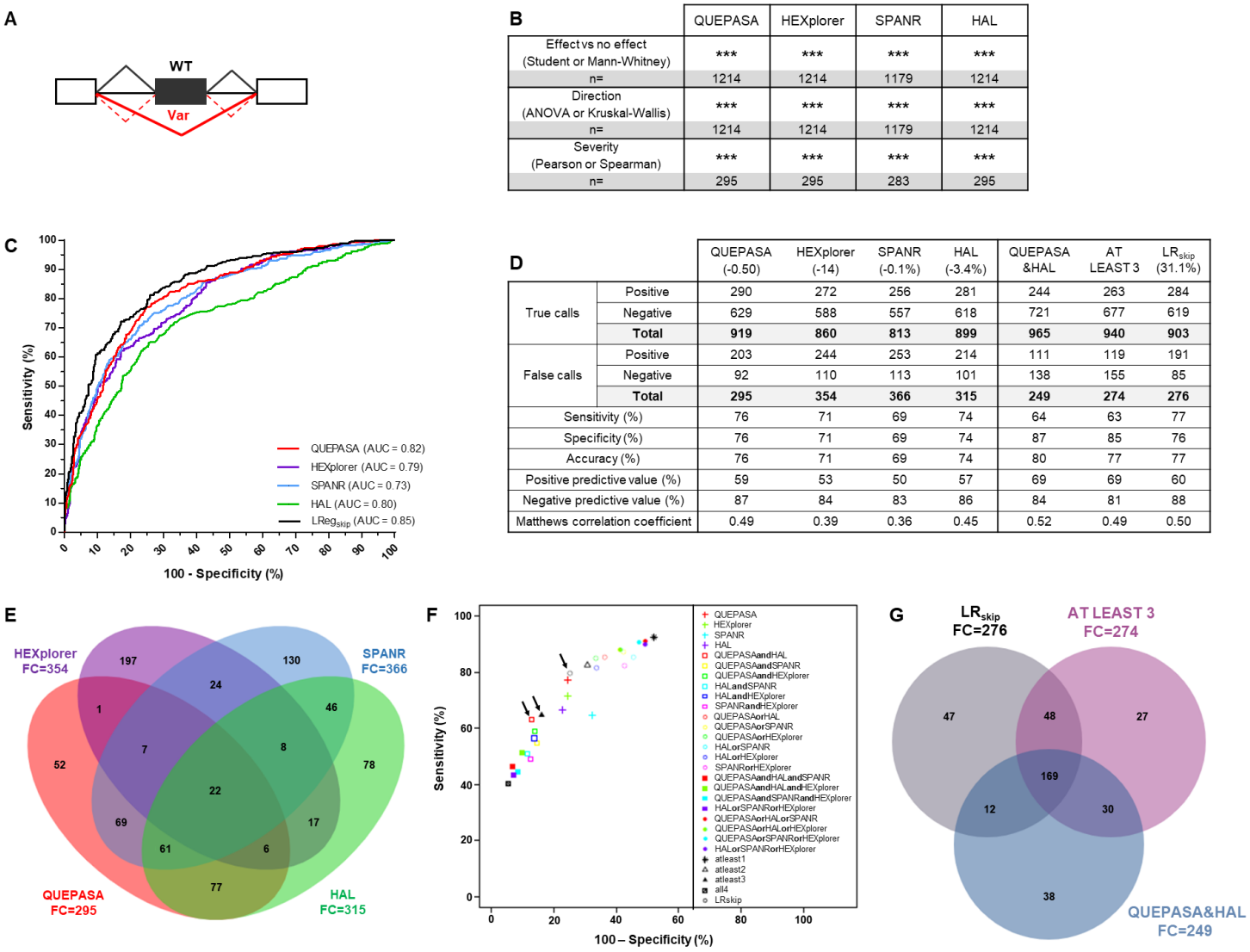
	QUEPASA	HEXplorer	SPANR	HAL
Effect vs no effect (t-test or Mann-Whitney)	***	***	***	***
n=	154	154	149	154
Direction (ANOVA or Kruskal-Wallis)	n/s	n/s	n/s	n/s
n=	154	154	149	154
Severity (Pearson or Spearman)	***	***	***	***
n=	124	124	119	124



E

		Soukarieh *et al.*, 2016			This study	
		QUEPASA (-0,50)	HEXplorer (-20)	SPANR (-5)	SPANR (-0,38)	HAL (-5)
True calls	Positives	40	38	9	38	45
	Negatives	84	77	95	77	75
	Total	**124**	**115**	**104**	**115**	**120**
False calls	Positives	20	27	3	27	29
	Negatives	10	12	41	12	5
	Total	**30**	**39**	**44**	**39**	**34**
Sensitivity (%)		80	76	18	76	90
Specificity (%)		81	74	97	74	72
Accuracy (%)		81	75	70	75	78

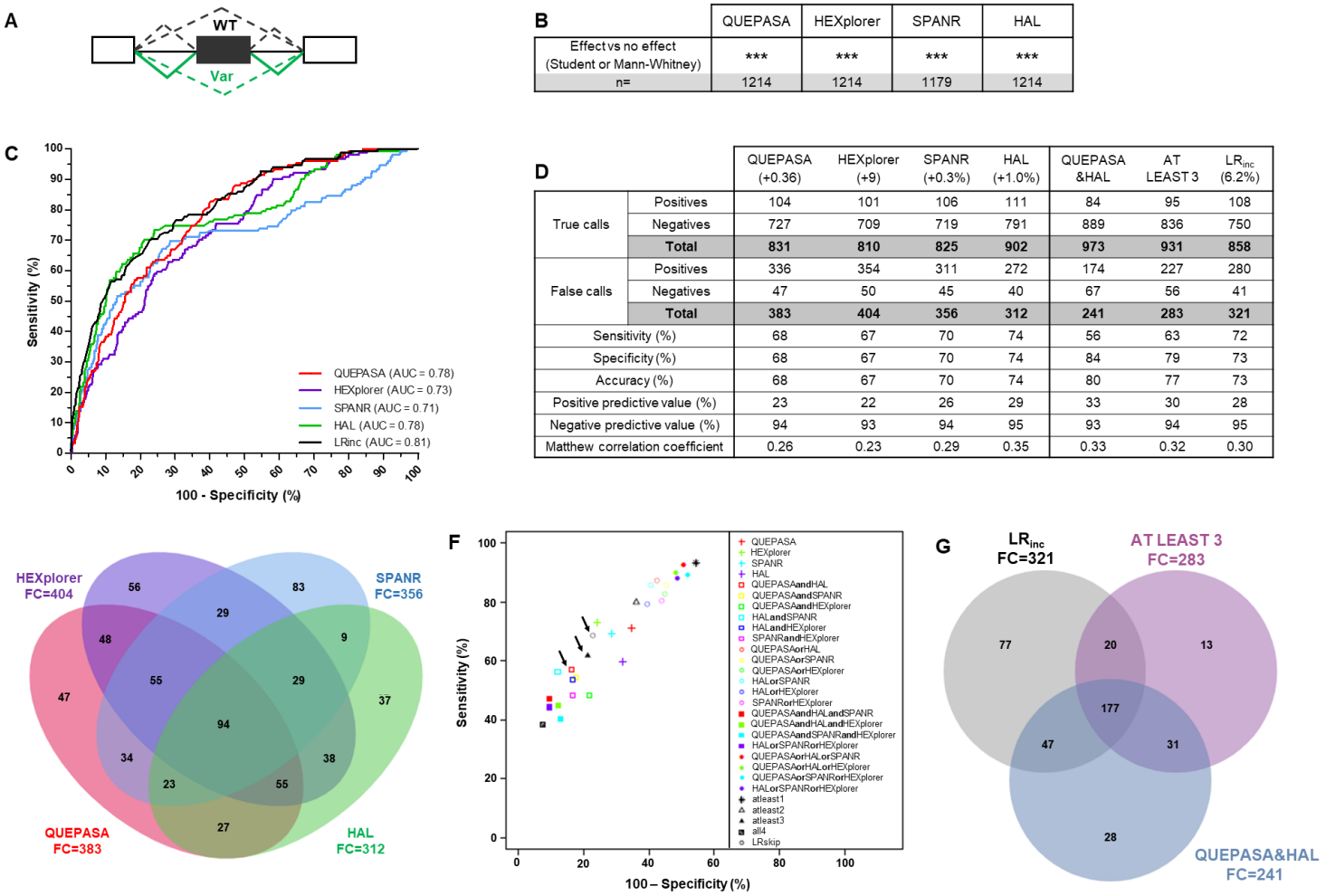
Figure 1. Small-scale assessment of the performance of SRE-dedicated bioinformatics predictions by using two previously described datasets.



B	QUEPASA	HEXplorer	SPANR	HAL
Effect vs no effect (Student or Mann-Whitney)	***	***	***	***
n=	1214	1214	1179	1214
Direction (ANOVA or Kruskal-Wallis)	***	***	***	***
n=	1214	1214	1179	1214
Severity (Pearson or Spearman)	***	***	***	***
n=	295	295	283	295

D		QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0.1%)	HAL (-3.4%)	QUEPASA &HAL	AT LEAST 3	LR_{skip} (31.1%)
True calls	Positive	290	272	256	281	244	263	284
	Negative	629	588	557	618	721	677	619
	Total	**919**	**860**	**813**	**899**	**965**	**940**	**903**
False calls	Positive	203	244	253	214	111	119	191
	Negative	92	110	113	101	138	155	85
	Total	**295**	**354**	**366**	**315**	**249**	**274**	**276**
Sensitivity (%)		76	71	69	74	64	63	77
Specificity (%)		76	71	69	74	87	85	76
Accuracy (%)		76	71	69	74	80	77	77
Positive predictive value (%)		59	53	50	57	69	69	60
Negative predictive value (%)		87	84	83	86	84	81	88
Matthews correlation coefficient		0.49	0.39	0.36	0.45	0.52	0.49	0.50

Figure 2. Large-scale evaluation of the performance of SRE-dedicated *in silico* tools in predicting variant-induced exon skipping by using an extensive training dataset.

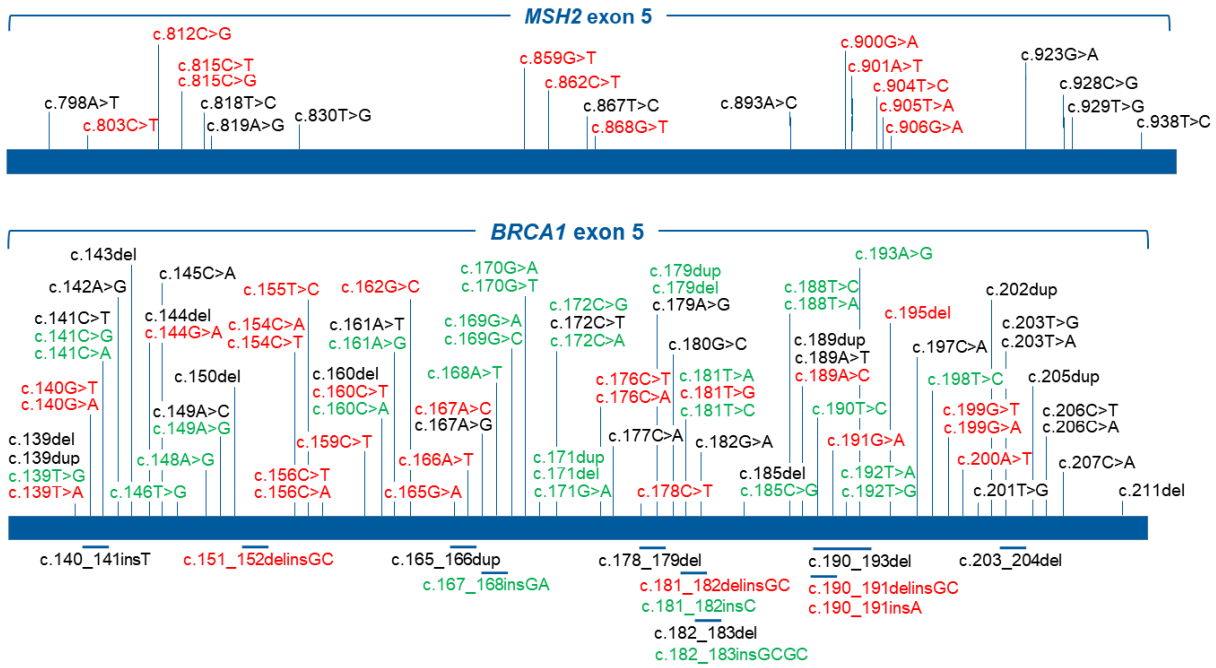


B	QUEPASA	HEXplorer	SPANR	HAL
Effect vs no effect (Student or Mann-Whitney)	***	***	***	***
n=	1214	1214	1179	1214

D		QUEPASA (+0.36)	HEXplorer (+9)	SPANR (+0.3%)	HAL (+1.0%)	QUEPASA &HAL	AT LEAST 3	LR_{inc} (6.2%)
True calls	Positives	104	101	106	111	84	95	108
	Negatives	727	709	719	791	889	836	750
	Total	**831**	**810**	**825**	**902**	**973**	**931**	**858**
False calls	Positives	336	354	311	272	174	227	280
	Negatives	47	50	45	40	67	56	41
	Total	**383**	**404**	**356**	**312**	**241**	**283**	**321**
Sensitivity (%)		68	67	70	74	56	63	72
Specificity (%)		68	67	70	74	84	79	73
Accuracy (%)		68	67	70	74	80	77	73
Positive predictive value (%)		23	22	26	29	33	30	28
Negative predictive value (%)		94	93	94	95	93	94	95
Matthew correlation coefficient		0.26	0.23	0.29	0.35	0.33	0.32	0.30

Figure 3. Large-scale evaluation of the performance of SRE-dedicated *in silico* tools in predicting variant-increased exon inclusion by using an extensive training dataset.

A



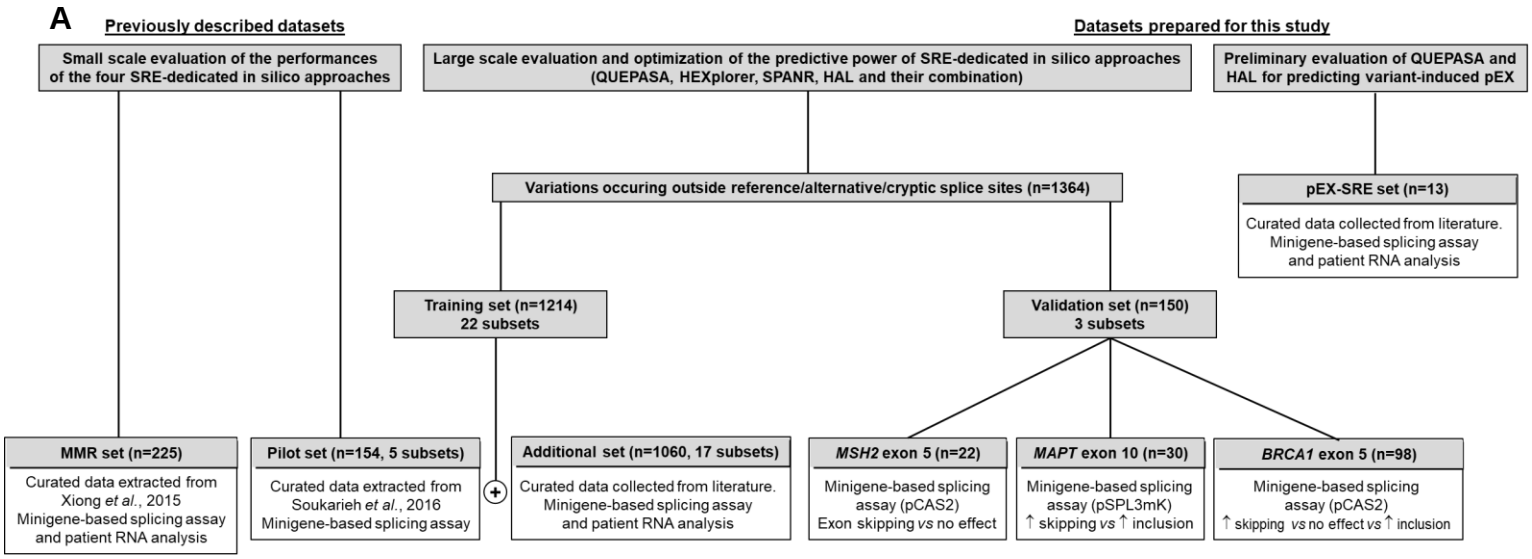
B

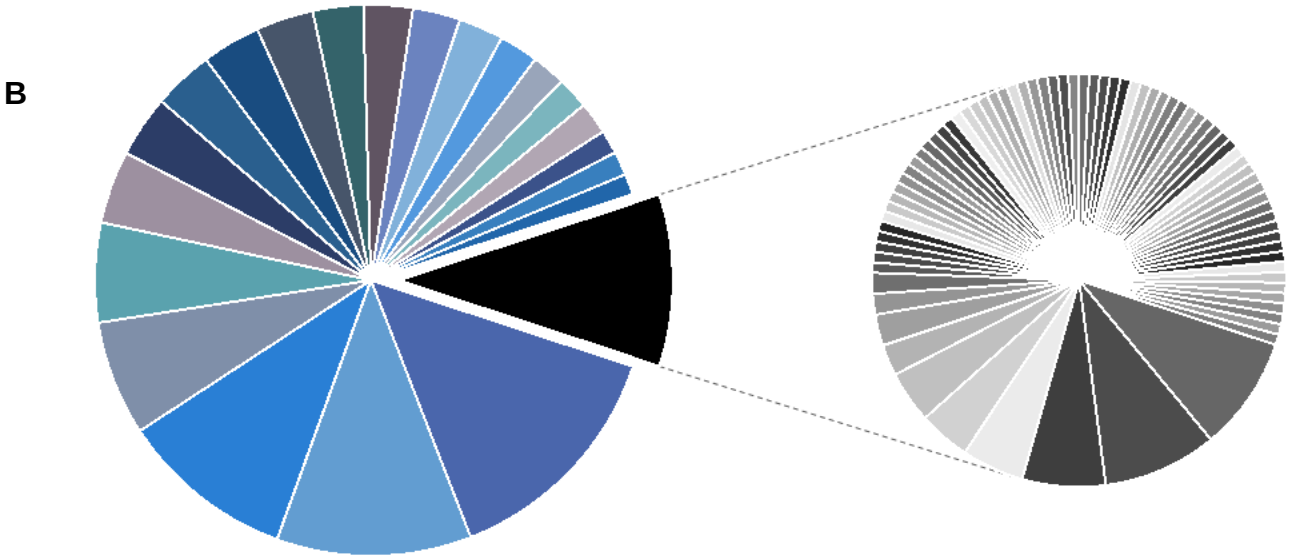
		QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0.1%)	HAL (-3.4%)	QUEPASA &HAL	AT LEAST 3	LR_{skip} (31.1%)
True calls	Positive	45	44	42	43	40	42	44
	Negative	86	73	44	79	92	86	84
	Total	131	117	86	122	132	128	128
False calls	Positive	9	22	25	16	3	9	9
	Negative	10	11	7	12	15	13	12
	Total	19	33	32	28	18	22	21
Sensitivity (%)		82	80	86	78	73	76	79
Specificity (%)		91	77	64	83	97	91	90
Accuracy (%)		87	78	73	81	88	85	86
Positive predictive value (%)		83	67	63	73	93	82	83
Negative predictive value (%)		90	87	85	87	86	87	88
Matthews correlation coefficient		0.73	0.55	0.49	0.61	0.74	0.68	0.70

C

		QUEPASA (+0,36)	HEXplorer (+9)	SPANR (+0,3%)	HAL (+1%)	QUEPASA &HAL	AT LEAST 3	LR_{inc} (6.3%)
True calls	Positive	39	40	30	43	37	40	38
	Negative	74	77	63	78	94	88	86
	Total	113	117	93	121	131	128	124
False calls	Positive	28	25	16	24	8	14	16
	Negative	9	8	9	5	11	8	9
	Total	37	33	25	29	19	22	26
Sensitivity (%)		81	83	77	90	77	83	81
Specificity (%)		73	75	80	76	92	86	84
Accuracy (%)		75	78	79	81	87	85	83
Positive predictive value (%)		58	62	65	64	82	74	70
Negative predictive value (%)		89	91	88	94	90	92	91
Matthews correlation coefficient		0.50	0.55	0.49	0.61	0.70	0.68	0.62

Figure 4. A prospective study identifies new disease-associated splicing mutations and validates SRE-dedicated *in silico* approaches for predicting variant-induced splicing alterations.

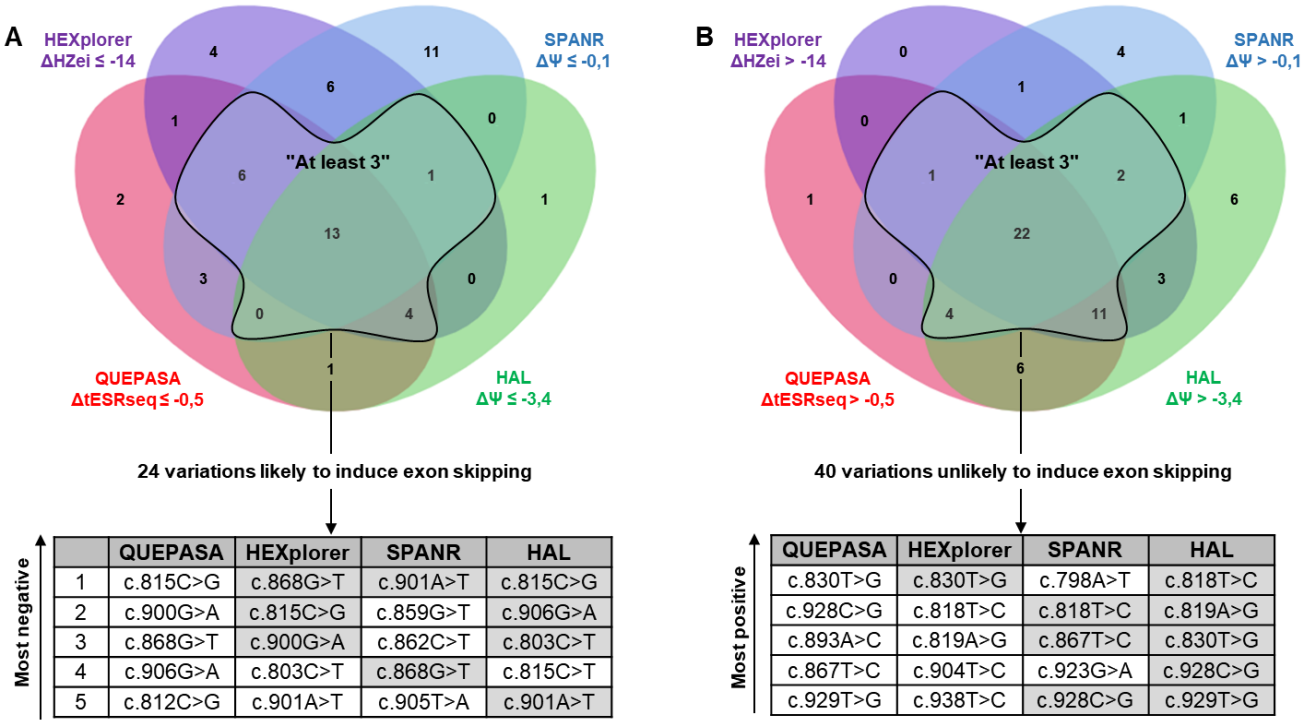




Splicing defects \ Type of data	Qualitative (number of variants)	Qualitative and quantitative (number of variants)
No effect *versus* **Increased exon skipping**	BRCA1 exon 18 (n=28) BRCA2 exon 18 (n=23) NF1 exon 37 (n=24)* ABCB11 (n=82) BRCA1 (n=15) BRCA2 (n=52) DYSF (n=23) MLH1 (n=71) MSH2 (n=31) Others (n=123)	BRCA2 exon 7 (n=34)* CFTR exon 12 (n=41)* SMN1 exon 7 (n=36)
No effect *versus* **Increased exon skipping** or **Increased exon inclusion**	BRCA1 exon 11 (n=125) FAS exon 6 (n=171) WT1 exon 5 (n=139)	BRCA1 exon 6 (n=42)* CFTR exon 9 (n=44) FIX exon 5 (n=17) MLH1 exon 10 (n=15)* NF1 exon 9 (n=35) SMN2 exon 7 (n=43)

* Subset included in the pilot dataset previously reported in Soukarieh *et al.*, 2016

Figure S1. Datasets used in this study for evaluating the predictive power of SRE-dedicated *in silico* approaches.



Most negative

	QUEPASA	HEXplorer	SPANR	HAL
1	c.815C>G	c.868G>T	c.901A>T	c.815C>G
2	c.900G>A	c.815C>G	c.859G>T	c.906G>A
3	c.868G>T	c.900G>A	c.862C>T	c.803C>T
4	c.906G>A	c.803C>T	c.868G>T	c.815C>T
5	c.812C>G	c.901A>T	c.905T>A	c.901A>T

Most positive

QUEPASA	HEXplorer	SPANR	HAL
c.830T>G	c.830T>G	c.798A>T	c.818T>C
c.928C>G	c.818T>C	c.818T>C	c.819A>G
c.893A>C	c.819A>G	c.867T>C	c.830T>G
c.867T>C	c.904T>C	c.923G>A	c.928C>G
c.929T>G	c.938T>C	c.928C>G	c.929T>G

Figure S2. Selection of 22 *MSH2* exon 5 variants for a prospective study and integration into the validation dataset.

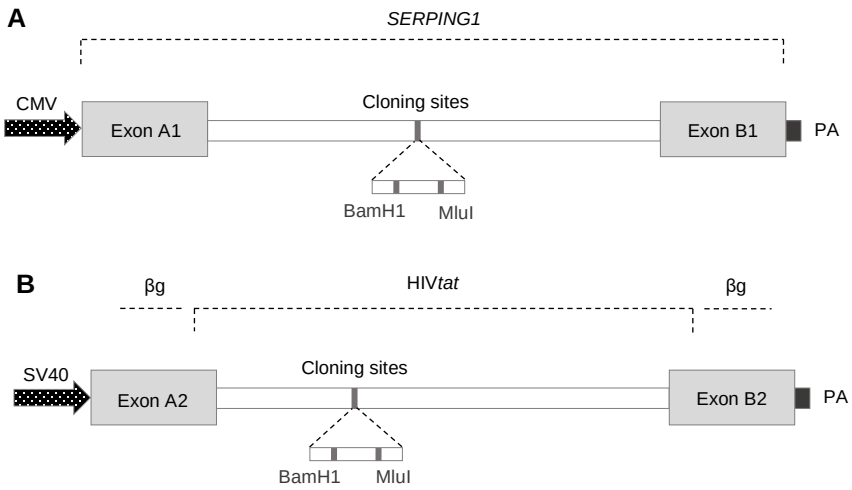


Figure S3. Structure of the pCAS2 and pSPL3mK constructs used in the cell-based minigene splicing assays.

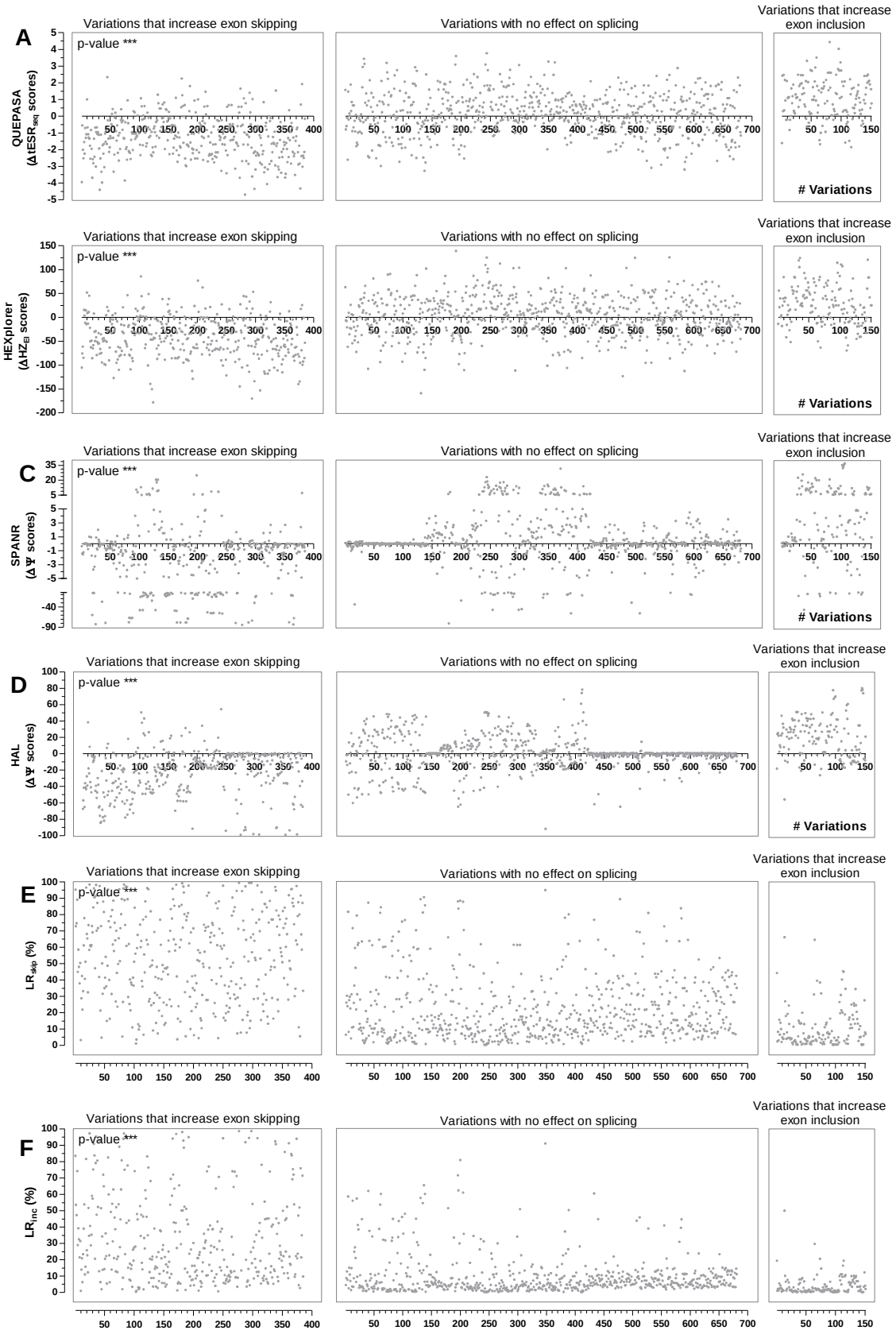


Figure S4. Comparison of the variant-associated splicing effects described in the training dataset with *in silico* data obtained with SRE-dedicated approaches.

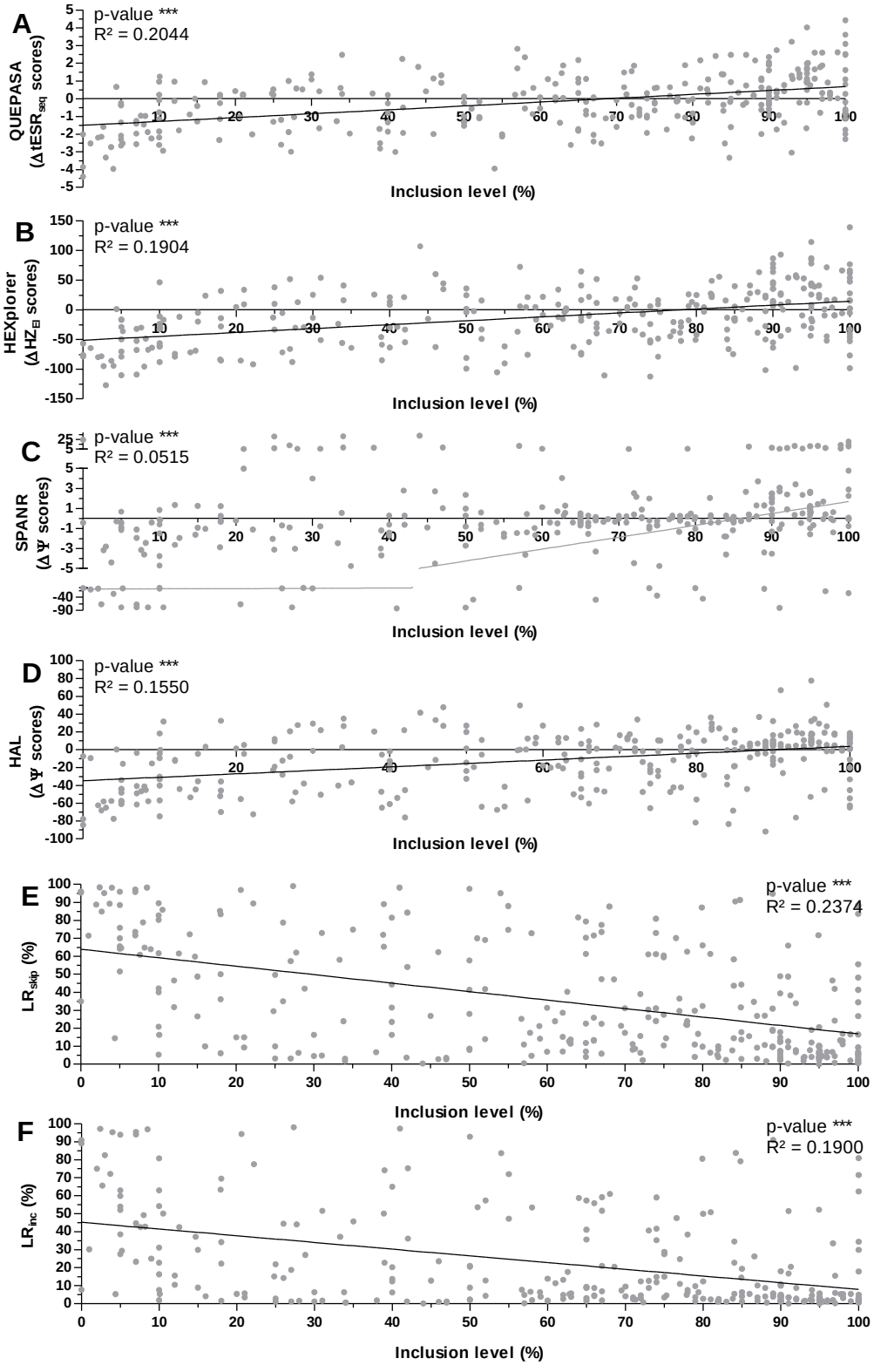


Figure S5. Correlation between variant-associated exon inclusion levels described in the training dataset and *in silico* data obtained with SRE-dedicated approaches.

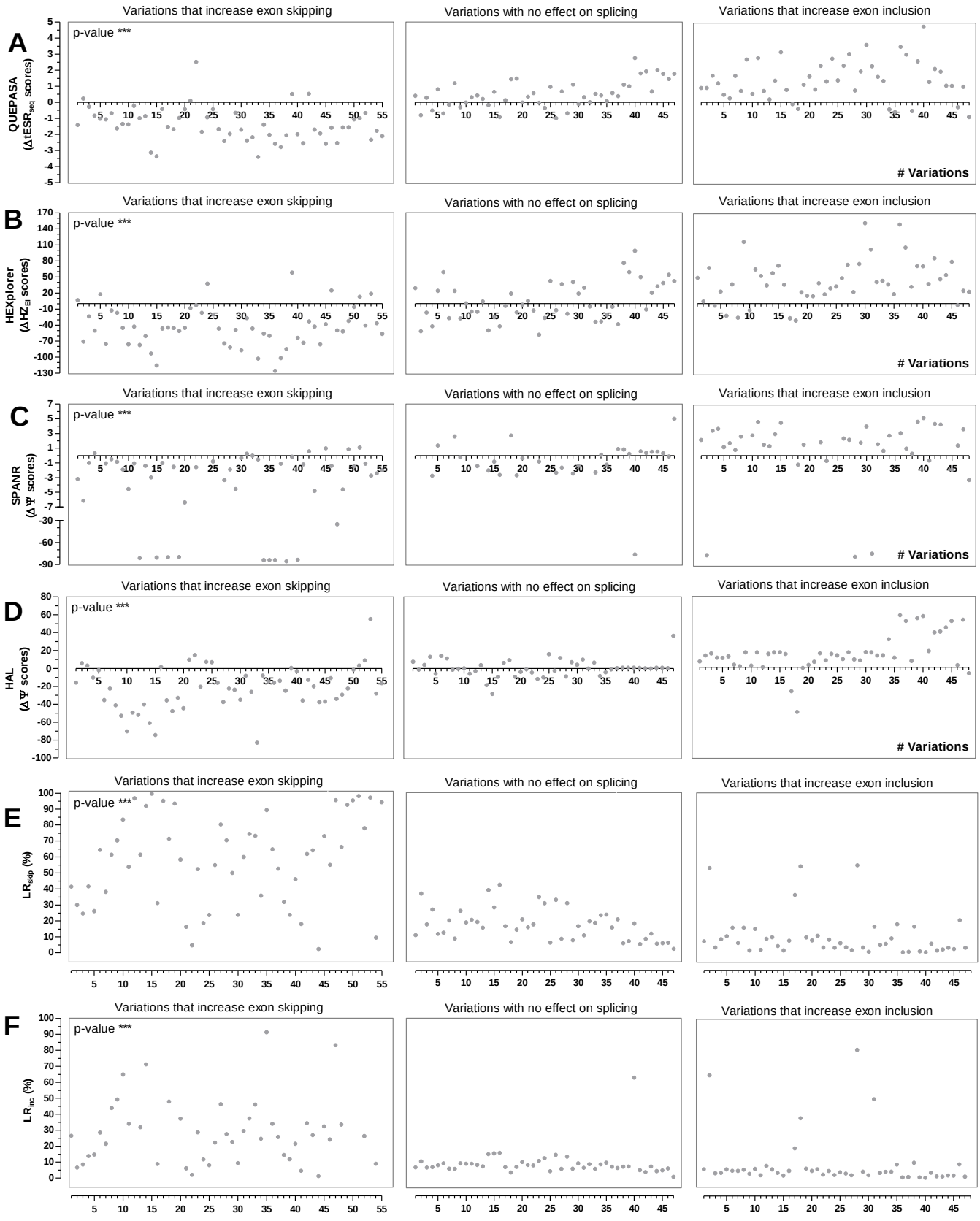


Figure S6. Comparison of the variant-associated splicing effects observed in the validation dataset with *in silico* data obtained with SRE-dedicated approaches.

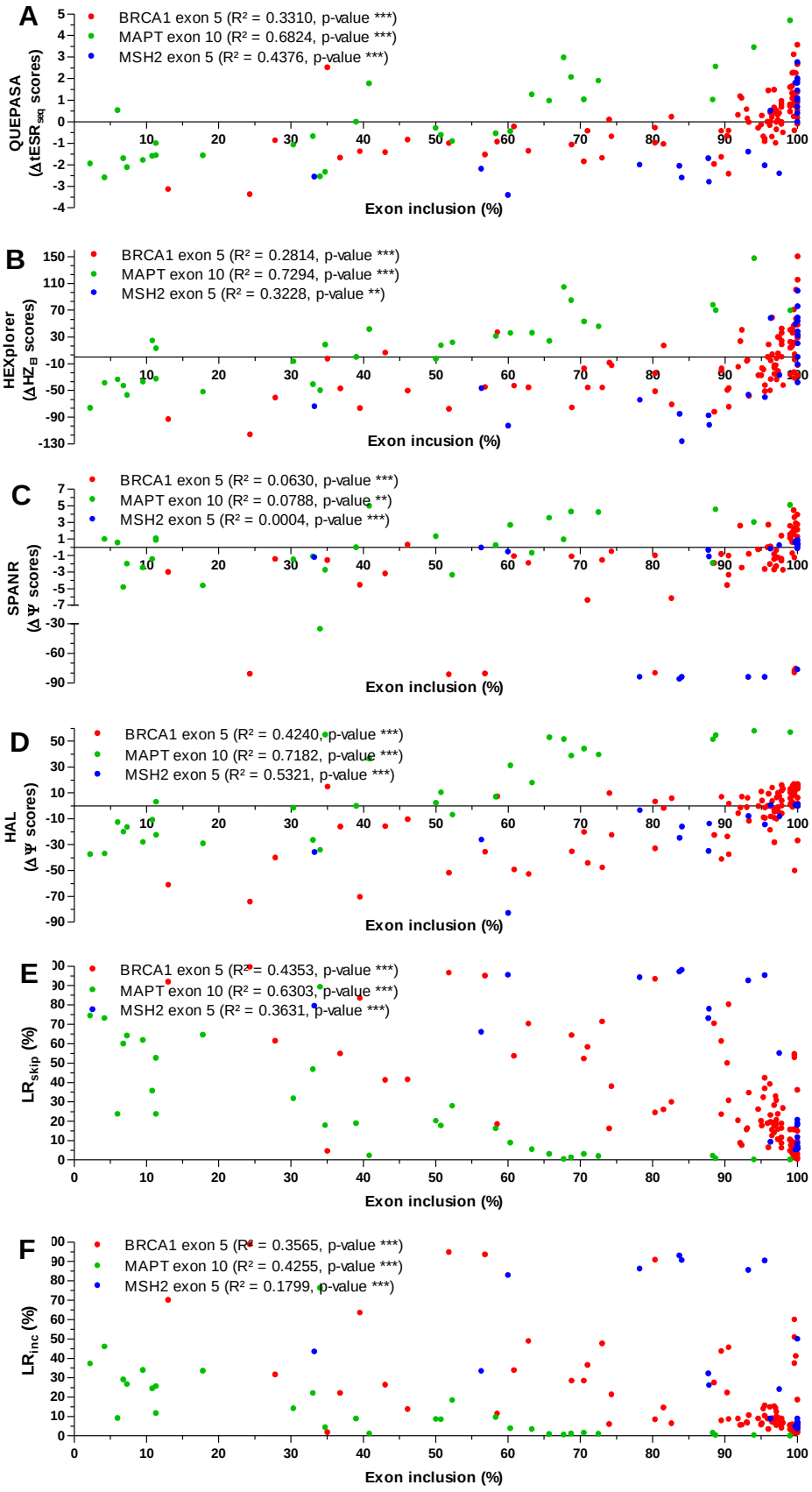
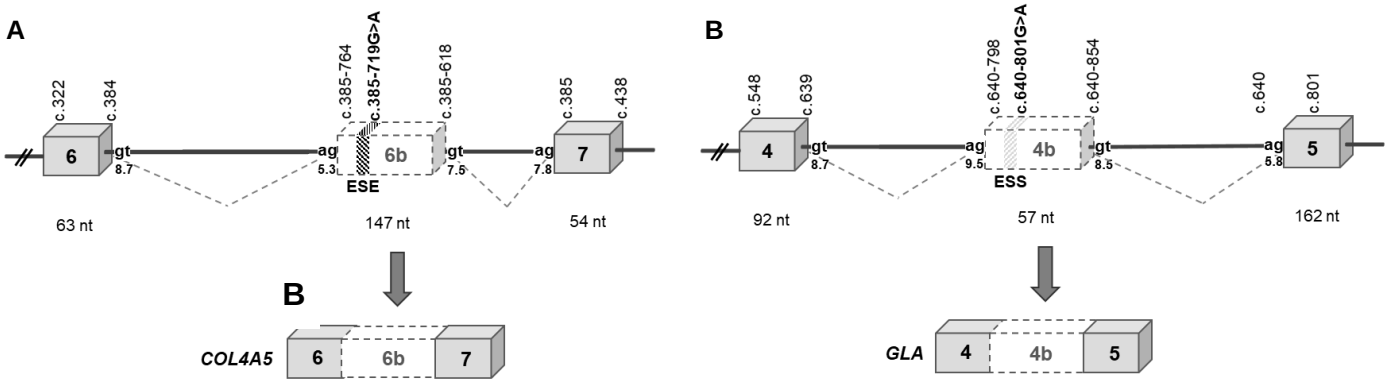


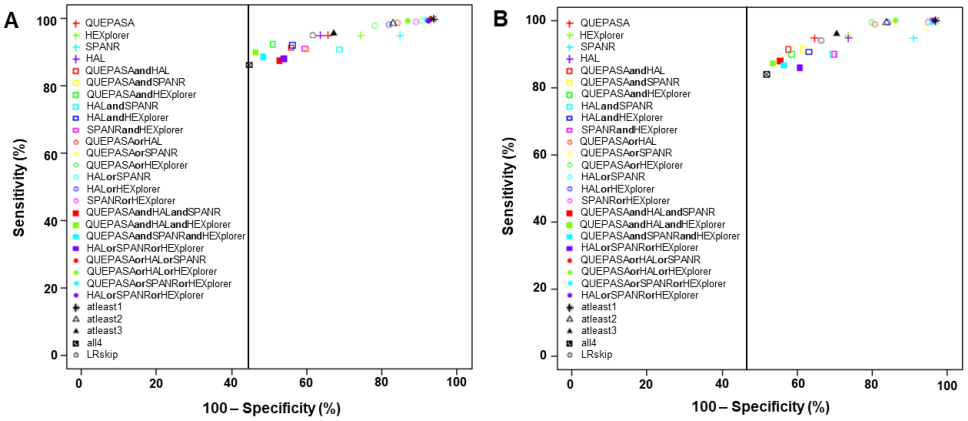
Figure S7. Correlation between variant-associated exon inclusion levels observed in the validation dataset and *in silico* data obtained with SRE-dedicated approaches.



C

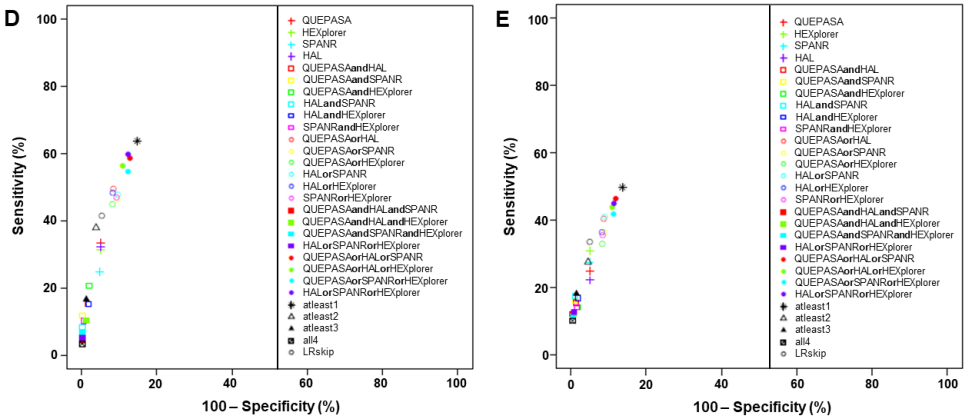
Variants				Pseudoexon (pEx)				SRE-dedicated bioinformatic predictions		References
Associated disease	Gene	Transcript (NM)	Nucleotide variation	Coordinates (start-end)	Size (bp)	Variant distance from pEX termini		QUEPASA (+0.36)	HEXplorer (+9)	
Stargardt disease	*ABCA4*	000350.2	c.4539+2001G>A	c.4539+1891 c.4540-2162	345	+111	-235	3,02	80,0	[1,2]
Stargardt disease	*ABCA4*	000350.2	c.4539+2028C>T	c.4539+1891 c.4540-2162	345	+138	-208	-0,43	2,1	[1,3]
Ataxia-telangiectasia	*ATM*	000051.3	c.2839-581_2839-578del	c.2839-593 c.2839-525	69	+12	-54	2,1	49,5	[4-8]
Cystic Fibrosis	*CFTR*	000492.3	c.870-1113_870-1110del	c.870-1162 c.870-1062	101	+50	-49	1,98	79,5	[9-10]
Alport Syndrome	*COL4A5*	033380.2	c.385-719G>A	c.385-764 c.385-618	147	+46	-102	2,56	56,8	[11]
Duchenne muscular dystrophy	*DMD*	004006.2	c.3603+2053G>C	c. 3603+2033 c. 3603+2112	80	+21	-60	0,38	9,5	[12]
Hemophilia A	*F8*	000132.3	c.2113+1152delA	c.2113+1144 c.2113+1256	113	+9	-104	2,63	36,6	[13]
Afibrinogenemia	*FGB*	005141.4	c.115-600A>G	c.115-622 c.115-573	50	+23	-28	2,49	109,4	[14-16]
Krabbe's disease	*GALC*	000153.3	c.621+770T>C	c.621+758 c.621+791	34	+13	-22	1,29	72,0	[17]
Fabry disease	*GLA*	000169.2	c.640-801G>A	c.640-854 c.640-798	57	+54	-4	-0,50	8,1	[18-19]
Homocystinuria	*MTRR*	024010.2	c.984+469T>C	c.984+447 c. 984+586	140	+23	-118	1,71	3,8	[20-21]
Neurofibromatosis type 1	*NF1*	000267.3	c.888+744A>G	c.888+710 c.888+784	75	+35	-41	-0,10	11,9	[22]
Propionicacidemia	*PCCA*	000282.3	c.1285-1416A>G	c. 1285-1441 c. 1285-1358	84	+26	-59	2,46	74,6	[23]

Figure S8. Prediction of pseudoexon inclusion triggered by deep intronic variants suspected to alter splicing regulatory elements.



C

		QUEPASA		HEXplorer		SPANR		HAL		QUEPASA &HAL		AT LEAST 3		LR_{skip} & LR_{inc}	
		+0.83	-0.95	+23.7	-42.1	+5.1%	-7.3%	12.7%	-16.7%					10.4%	17.9%
True calls	Positive	363	144	363	144	351	142	363	144	349	138	364	147	350	144
	Negative	287	369	296	274	121	92	226	279	375	445	284	312	324	305
	Total	650	513	659	418	472	234	589	423	724	583	648	459	674	449
False calls	Positive	545	694	536	789	689	938	606	784	457	618	548	751	486	724
	Negative	19	7	19	7	18	7	19	19	33	13	18	4	19	7
	Total	564	1703	555	796	707	945	625	803	490	631	566	755	505	731
Sensitivity (%)		95	95	95	95	95	95	95	95	91	91	95	97	95	95
Specificity (%)		35	35	36	26	15	9	27	1	45	42	34	29	40	30
Accuracy (%)		54	42	54	34	40	20	49	35	60	48	53	38	57	38
Positive predictive value (%)		40	17	40	15	34	13	37	16	43	18	40	16	42	17
Negative predictive value (%)		94	98	94	98	87	93	92	98	92	97	94	99	94	98
Matthews correlation coefficient		0.32	0.21	0.32	0.17	0.15	0.05	0.26	0.17	0.36	0.23	0.32	0.20	0.36	0.19



F

		QUEPASA		HEXplorer		SPANR		HAL		QUEPASA &HAL		AT LEAST 3		LR_{skip} & LR_{inc}	
		-1.87	+1.95	-63.0	+60.0	-5.1%	+7.1%	-35.1%	+32.4%					63.7%	1.0%
True calls	Positive	126	37	122	33	91	41	123	47	64	23	65	27	149	50
	Negative	792	1010	791	1010	771	979	791	1010	819	1049	822	1052	771	980
	Total	918	1047	913	1043	862	1020	914	1057	883	1072	887	1079	920	1030
False calls	Positive	40	53	41	53	39	51	41	53	13	14	10	11	39	50
	Negative	256	114	260	118	278	108	259	104	318	128	317	124	220	99
	Total	296	167	301	171	317	159	300	157	331	142	327	135	259	154
Sensitivity (%)		33	25	32	22	25	28	32	31	17	15	17	78	40	34
Specificity (%)		95	95	95	95	95	95	95	95	98	99	99	99	95	95
Accuracy (%)		76	86	75	86	73	87	75	87	73	88	73	89	79	87
Positive predictive value (%)		76	41	75	38	70	45	75	47	83	62	87	71	79	50
Negative predictive value (%)		76	90	75	90	74	90	75	91	72	89	72	89	78	91
Matthews correlation coefficient		0.38	0.25	0.37	0.22	0.29	0.28	0.37	0.32	0.29	0.27	0.31	0.32	0.45	034

Figure S9. Large-scale evaluation of the performance of SRE-dedicated *in silico* tools in predicting variant-induced splicing alterations at 95% specificity or sensitivity by using an extensive training dataset.

	Approaches	Scores (Thresholds)	*In silico* tools	Input	Output	Application	Variant numbers	Ref
QUEPASA	Experimental assessment of the ESR properties of all possible 6-nucleotide motifs (4096 hexamers) using splicing reporter minigene assays	$\Delta tESR_{seq}$ (Skip = -0.5) (Inc = +0.36)	Alamut Batch (https://www.interactive-biosoftware.com/alamut-batch/) *NB: Takes into account all possible hexamers including the ones overlapping with intronic positions*	VCF files	Annotations for NGS analysis in tab delaminated format (variant description, database, splicing and missense predictions)	Any types of variants	Illimited	[1,2]
			HExoSplice (http://bioinfo.univ-rouen.fr/HExoSplice_submit/) *NB: By default takes into account exonic hexamers only (but can be deflected to analyze intronic positions)*	WT sequence in fasta format as well as variant identity	CSV file with WT and mutant tESRseq scores and $\Delta tESR_{seq}$ scores Distribution of WT ESRseq scores within the exon	Single base substitutions	Illimited per exon	
HEXplorer	Statistical approach derived from a RESCUE-type analysis that computed the relative distribution of hexamer motifs in exons and introns.	ΔHZ_{EI} (Skip = -14) (Inc = +9)	Alamut Batch (https://www.interactive-biosoftware.com/alamut-batch/)	VCF files	Annotations for NGS analysis in tab delaminated format (variant description, database census, splicing and missense predictions)	Any types of variants	Illimited	[3]
			HEXplorer Score (https://www2.hhu.de/rna/html/hexplorer_score.php)	ASCII text file (line triples in FASTA like format) containing variant name, WT and mutant sequence	HEXplorer scores (ΔHZ_{EI}) HEXplorer score profiles of the exon	Exonic single base substitutions	One by one	
			Makro_HEXplorer (https://academic.oup.com/nar/article/42/16/10681/2903056)	Excel files with WT and mutant sequence as well as position and nature of nucleotide change	HEXplorer scores (ΔHZ_{EI}) HEXplorer score profiles of the exon	Single base substitutions	Illimited per exon	
SPANR	Machine learning approach correlating 1393 characteristics extracted from the DNA sequence located in the environment of the mutation with splicing data obtained by RNAseq from 16 different healthy human tissues	$\Delta\Psi$ (%) (Skip = -0.1) (Inc = +0.3)	SPANR alpha http://tools.genes.toronto.edu/	VCF files including chromosome number, genomic position, variant name, reference and mutated allele	The maximum mutation-induced change in PSI across 16 tissues, how this value compares to those for common SNPs in the form of percentiles, and the predicted average wildtype PSI in the 16 tissues in tab delaminated format	Single base substitutions outside terminal exons or very large exon. Intronic variants should be located up to 300nt from splice junctions	40 variants at time (2 resquest per hour)	[4]
HAL	Combination of synthetic biology and machine learning approaches using splicing profiles generated from a library of nearly 2 million spliced synthetic minigenes	$\Delta\Psi$ (%) (Skip = -3.4) (Inc = +1.0)	HAL Splice Predictions http://splicing.cs.washington.edu/SE	Txt files with tab deliminaters including variant name, WT and variant sequence (exon + 6 nt of the downstream intron) and inclusion level for the WT sequence (between 0-100)	The predicted PSI for the WT as well as the change in PSI ($\Delta\Psi$) in tab delaminated format	Any types of exonic variations	Illimited One by one	[5]

Table S1. Overview of the four SRE-dedicated *in silico* approaches evaluated in this study.

Purpose	Name[1]	Sequence (5'-3')[2]
Site-directed mutagenesis	MSH2 Ex5 c.798AT-R	GTGATGAAACAGCAACCTAA
	MSH2 Ex5 c.803CT-F	GTTGCAGTTTTATCACTGTCTG
	MSH2 Ex5 c.812CG-R	GATTACCGCACACAGTGATGA
	MSH2 Ex5 c.815CT-F	TCACTGTCTGTGGTAATCAAG
	MSH2 Ex5 c.818TC-R	CTAAAAACTTGATTGCCGCAGACAG
	MSH2 Ex5 c.819AG-F	CTGTCTGCGGTGATCAAGTTTTTAG
	MSH2 Ex5 c.830TG-R	GATAAGAGTTCTCAAAACTTGA
	MSH2 Ex5 c.859GT-F	GATTCCAACTTTTGACAGTTTGAAC
	MSH2 Ex5 c.862CT-F	CAACTTTGGATAGTTTGAACTG
	MSH2 Ex5 c.867TC-R	GTAGTCAGTTCGAACTGTCCAA
	MSH2 Ex5 c.893AC-F	GACTTCAGCCCGTATATGAAATTG
	MSH2 Ex5 c.900GA-R	CAATATCCAATTTTATATACTGGC
	MSH2 Ex5 c.901AT-F	GCCAGTATATGTAATTGGATATTG
	MSH2 Ex5 c.904TC-R	CTGCAATATCCAGTTTCATATAC
	MSH2 Ex5 c.905TA-F	GTATATGAAATAGGATATTGCAG
	MSH2 Ex5 c.906GA-R	CTGCAATATCTAATTTCATATAC
	MSH2 Ex5 c.923GA-F	GCAGCAGTCAAAGCCCTTAAC
	MSH2 Ex5 c.928CG-R	GAAAAAGGTTAACGGCTCTGACTG
	MSH2 Ex5 c.929TG-F	CAGTCAGAGCCCGTAACCTTTTTC
	MSH2 Ex5 c.938TC-R	TTTTACCTGAGAAAGGTTAAG
	BRCA1 Ex5 c.139TG-F	TTATAGATTTGGCATGCTGAAAC
	BRCA1 Ex5 c.139TA-F	TTATAGATTTAGCATGCTGAAAC
	BRCA1 Ex5 c.139del-F	TTATAGATTTGCATGCTGAAAC
	BRCA1 Ex5 c.139dup-F	TTATAGATTTTTGCATGCTGAAAC
	BRCA1 Ex5 c.140GT-F	TTATAGATTTTTCATGCTGAAAC
	BRCA1 Ex5 c.140GA-F	TTATAGATTTTACATGCTGAAAC
	BRCA1 Ex5 c.140_141insT-F	TTATAGATTTTGTCATGCTGAAAC
	BRCA1 Ex5 c.141CA-F	ATAGATTTTGAATGCTGAAACTTC
	BRCA1 Ex5 c.141CG-F	ATAGATTTTGGATGCTGAAACTTC
	BRCA1 Ex5 c.141CT-F	ATAGATTTTGTATGCTGAAACTTC
	BRCA1 Ex5 c.142AG-F	ATAGATTTTGCGTGCTGAAACTTC
	BRCA1 Ex5 c.143del-F	ATAGATTTTGCAGCTGAAACTTC
	BRCA1 Ex5 c.144GA-F	GATTTTGCATACTGAAACTTCTC
	BRCA1 Ex5 c.144del-F	GATTTTGCATCTGAAACTTCTC
	BRCA1 Ex5 c.145CA-F	GATTTTGCATGATGAAACTTCTC
	BRCA1 Ex5 c.146TG-F	TTTGCATGCGGAAACTTCTCAAC
	BRCA1 Ex5 c.148AG-F	TTTGCATGCTGGAACTTCTCAAC
	BRCA1 Ex5 c.149AG-F	TTTGCATGCTGAGACTTCTCAAC
	BRCA1 Ex5 c.149AC-F	TTTGCATGCTGACACTTCTCAAC
	BRCA1 Ex5 c.150del-F	GCATGCTGAACTTCTCAACCAG
	BRCA1 Ex5 c.151_152delinsGC-F	GCATGCTGAAAGCTCTCAACCAG
	BRCA1 Ex5 c.154CT-F	TGCTGAAACTTTTCAACCAGAAG
	BRCA1 Ex5 c.154CA-F	TGCTGAAACTTATCAACCAGAAG
	BRCA1 Ex5 c.155TC-F	GCTGAAACTTCCCAACCAGAAGA
	BRCA1 Ex5 c.156CA-F	GCTGAAACTTCTAAACCAGAAGA
	BRCA1 Ex5 c.156CT-F	GCTGAAACTTCTTAACCAGAAGA
	BRCA1 Ex5 c.159CT-F	AACTTCTCAATCAGAAGAAAGGG
	BRCA1 Ex5 c.160CT-F	AACTTCTCAACTAGAAGAAAGGG
	BRCA1 Ex5 c.160CA-F	AACTTCTCAACAAGAAGAAAGGG
	BRCA1 Ex5 c.160del-F	AACTTCTCAACAGAAGAAAGGG
	BRCA1 Ex5 c.161AG-F	ACTTCTCAACCGGAAGAAAGGG
	BRCA1 Ex5 c.161AT-F	ACTTCTCAACCTGAAGAAAGGG
	BRCA1 Ex5 c.162GC-F	CTCAACCACAAGAAAGGGCCTTC

	BRCA1 Ex5 c.165GA-F	CTCAACCAGAAAAAAGGGCCTTC
	BRCA1 Ex5 c.165_166dup-F	CTCAACCAGAAGAGAAAGGGCCTTC
	BRCA1 Ex5 c.166AT-F	CAACCAGAAGTAAGGGCCTTCAC
	BRCA1 Ex5 c.167AG-F	CAACCAGAAGAGAGGGCCTTCAC
	BRCA1 Ex5 c.167AC-F	CAACCAGAAGACAGGGCCTTCAC
	BRCA1 Ex5 c.167_168insGA-F	CAACCAGAAGAAGAAGGGCCTTCAC
	BRCA1 Ex5 c.168AT-F	ACCAGAAGAATGGGCCTTCACAG
	BRCA1 Ex5 c.169GC-F	ACCAGAAGAAACGGCCTTCACAG
	BRCA1 Ex5 c.169GA-F	CAGAAGAAAAGGCCTTCACAGTG
	BRCA1 Ex5 c.170GT-F	CAGAAGAAAGTGCCTTCACAGTG
	BRCA1 Ex5 c.170GA-F	CAGAAGAAAGAGCCTTCACAGTG
	BRCA1 Ex5 c.171GA-F	CAGAAGAAAGGACCTTCACAGTG
	BRCA1 Ex5 c.171del-F	CAGAAGAAAGGCCTTCACAGTG
	BRCA1 Ex5 c.171dup-F	CAGAAGAAAGGGGCCTTCACAGTG
	BRCA1 Ex5 c.172CT-F	GAAGAAAGGGTCTTCACAGTGTC
	BRCA1 Ex5 c.172CG-F	GAAGAAAGGGGCTTCACAGTGTC
	BRCA1 Ex5 c.172CA-F	GAAGAAAGGGACTTCACAGTGTC
	BRCA1 Ex5 c.176CA-F	GAAAGGGCCTTAACAGTGTCCTT
	BRCA1 Ex5 c.176CT-F	GAAAGGGCCTTTACAGTGTCCTT
	BRCA1 Ex5 c.177CA-F	AACTTCTCAACAGAAGAAAGGG
	BRCA1 Ex5 c.178CT-R	TAAAGGACACTATGAAGGCCCTT
	BRCA1 Ex5 c.178_179del-R	ATAAAGGACACTGAAGGCCCTTTC
	BRCA1 Ex5 c.179AG-R	TAAAGGACACCGTGAAGGCCCTT
	BRCA1 Ex5 c.179del-R	TAAAGGACACGTGAAGGCCCTT
	BRCA1 Ex5 c.179dup-R	TAAAGGACACTTGTGAAGGCCCTT
	BRCA1 Ex5 c.180GC-R	CATAAAGGACAGTGTGAAGG
	BRCA1 Ex5 c.181TG-R	TACATAAAGGACCCTGTGAAGGC
	BRCA1 Ex5 c.181TC-R	TACATAAAGGACGCTGTGAAGGC
	BRCA1 Ex5 c.181TA-R	TACATAAAGGACTCTGTGAAGGC
	BRCA1 Ex5 c.182GA-R	TACATAAAGGATACTGTGAAGGC
	BRCA1 Ex5 c.181_182insC-R	TACATAAAGGACGACTGTGAAGGC
	BRCA1 Ex5 c.181_182delinsGC-R	TACATAAAGGAGCCTGTGAAGGC
	BRCA1 Ex5 c.182_183del-R	CTTACATAAAGGACTGTGAAGGC
	BRCA1 Ex5 c.182_183insGCGC-R	CTTACATAAAGGAGCGCCACTGTGAAGGC
	BRCA1 Ex5 c.185CG-R	CTTACATAAACGACACTGTGAAG
	BRCA1 Ex5 c.185del-R	CTTACATAAAGACACTGTGAAG
	BRCA1 Ex5 c.188TC-R	ATTCTTACATGAAGGACACTGTG
	BRCA1 Ex5 c.188TA-R	ATTCTTACATTAAGGACACTGTG
	BRCA1 Ex5 c.189AT-R	ATTCTTACAAAAAGGACACTGTG
	BRCA1 Ex5 c.189AC-R	ATTCTTACAGAAAGGACACTGTG
	BRCA1 Ex5 c.189dup-R	ATTCTTACATTAAAGGACACTGTG
	BRCA1 Ex5 c.190TC-R	TCATTCTTACGTAAAGGACACTG
	BRCA1 Ex5 c.190_191insA-R	TCATTCTTACTATAAAGGACACTG
	BRCA1 Ex5 c.190_191delinsGC-R	TCATTCTTAGCTAAAGGACACTG
	BRCA1 Ex5 c.190_193del-R	TATATCATTCTTAAAGGACACTG
	BRCA1 Ex5 c.191GA-R	TATCATTCTTATATAAAGGACAC
	BRCA1 Ex5 c.192TG-R	TATCATTCTTCCATAAAGGACAC
	BRCA1 Ex5 c.192TA-R	TATCATTCTTTCATAAAGGACAC
	BRCA1 Ex5 c.193AG-R	TATATCATTCTCACATAAAGGAC
	BRCA1 Ex5 c.195del-R	TTATATCATTTTACATAAAGGAC
	BRCA1 Ex5 c.197AG-R	GGTTATATCACTCTTACATAAAG
	BRCA1 Ex5 c.198TC-R	GGTTATATCGTTCTTACATAAAG
	BRCA1 Ex5 c.199GT-R	GGTTATATAATTCTTACATAAAG
	BRCA1 Ex5 c.199GA-R	GGTTATATTATTCTTACATAAAG

	BRCA1 Ex5 c.200AT-R	CTTTTGGTTATAACATTCTTAC
	BRCA1 Ex5 c.201TG-R	CTTTTGGTTATCTCATTCTTAC
	BRCA1 Ex5 c.202dup-R	CTTTTGGTTATTATCATTCTTAC
	BRCA1 Ex5 c.203TG-R	CCTTTTGGTTCTATCATTCTTAC
	BRCA1 Ex5 c.203TA-R	CCTTTTGGTTTTATCATTCTTAC
	BRCA1 Ex5 c.203_204del-R	TACCTTTTGGTTATCATTCTTAC
	BRCA1 Ex5 c.205dup-R	TATACCTTTTGGTTTTATATCATTC
	BRCA1 Ex5 c.206CA-R	TATACCTTTTGTTTATATCATTC
	BRCA1 Ex5 c.206CT-R	TATACCTTTTGATTATATCATTC
	BRCA1 Ex5 c.207CA-R	TATACCTTTTTGTTATATCATTC
	BRCA1 Ex5 c.211del-R	ATTATATACCTTTGGTTATATC
	MAPT Ex10 c.825GA-F	GCTACCAAAGGTACAGATAATTAAT
	MAPT Ex10 c.832AC-R	GCTTCTTATTAAGTATCTGCACCTT
	MAPT Ex10 c.836del-F	TGCAGATAATTATAAGAAGCTGGATC
	MAPT Ex10 c.837TG-R	ATCCAGCTTCTTCTTAATTATCTGC
	MAPT Ex10 c.837TC-F	GCAGATAATTAACAAGAAGCTGGAT
	MAPT Ex10 c.839AG-R	AGATCCAGCTTCCTATTAATTATCT
	MAPT Ex10 c.841_843del-F	GATAATTAATAAGCTGGATCTTAGC
	MAPT Ex10 c.849TG-R	GACGTTGCTAAGCTCCAGCTTCTTA
	MAPT Ex10 c.851TG-F	GAAGCTGGATCGTAGCAACGTCCAG
	MAPT Ex10 c.852TC-R	CTGGACGTTGCTGAGATCCAGCTTC
	MAPT Ex10 c.853AC-F	AGCTGGATCTTCGCAACGTCCAGTC
	MAPT Ex10 c.858CT-R	CTTGGACTGGACATTGCTAAGATCC
	MAPT Ex10 c.859GA-F	GATCTTAGCAACATCCAGTCCAAGT
	MAPT Ex10 c.862CT-R	CACACTTGGACTAGACGTTGCTAAG
	MAPT Ex10 c.871TC-F	GTCCAGTCCAAGCGTGGCTCAAAGG
	MAPT Ex10 c.878CT-R	GATATTATCCTTTAAGCCACACTTG
	MAPT Ex10 c.886AG-F	GGCTCAAAGGATGATATCAAACACG
	MAPT Ex10 c.886AC-R	CGTGTTTGATATGATCCTTTGAGCC
	MAPT Ex10 c.887_889del-F	GCTCAAAGGATATCAAACACGTCC
	MAPT Ex10 c.888TC-R	GACGTGTTTGATGTTATCCTTTGAG
	MAPT Ex10 c.897CT-F	TAATATCAAACATGTCCCGGGAGGC
	MAPT Ex10 c.898GA-R	CTCCCGGGATGTGTTTGATATTATC
	MAPT Ex10 c.901CT-F	TCAAACACGTCTCGGGAGGCGGCAG
	MAPT Ex10 c.901CA-R	TGCCGCCTCCCGTGACGTGTTTGAT
	MAPT Ex10 c.902CT-F	TCAAACACGTCCTGGGAGGCGGCAG
	MAPT Ex10 c.903GA-R	ACTGCCGCCTCCTGGGACGTGTTTG
	MAPT Ex10 c.908GT-F	ACGTCCCGGGAGTCGGCAGTGTGAG
	MAPT Ex10 c.909CT-R	ACTCACACTGCCACCTCCCGGGACG
	MAPT Ex10 c.910GA-F	GTCCCGGGAGGCAGCAGTGTGAGTA
	MAPT Ex10 c.912CT-R	GTACTCACACTACCGCCTCCCGGGAC
PCR (cloning, minigene preparation)	pCAS2_MSH2 Ex5 InFus BglII-F	AAGAAGTGCAGGATCTTTTTGGCTATTCTAAATAATGCTGCAAT
	pCAS2_MSH2 Ex5 InFus Mlu-R	TCAAAACAAGACGCGTAAAAAGTGGAGTGGAGGAGGG
	pCAS2_BRCA1 Ex5 InFus Bam-F	AAGAAGTGCAGGATCCGGAAACTATTGCTTGTAATTCACC
	pCAS2_BRCA1 Ex5 InFus Mlu-R	TCAAAACAAGACGCGTAGATGTCCCATAAAACTTTCAGG
	pSPL3_MAPT Ex10 InFus Bam-F	GAGCGGCCGCTGCAGGATCCGACTCAACCTCCCGTCACTC
	pSPL3_MAPT Ex10 InFus Mlu-R	GTACGGGATCACGCGTGGAACAGTGGACCGTGTGG
Sequencing of minigene inserts	pCAS-Seq-F	GGGTCAATAGCAGTGAGAGG
	pCAS-Seq-R	GCTCCATTTCACAGGTAGAGA
	pspl3-937F	CCTTGGGATGTTGATGATCTG
	pspl3-1191R	ACTTCTTGTGGGTTGGGGTC
RT-PCR and/or sequencing of	pCAS-KO1-F	TGACGTCGCCGCCCATCAC
	6FAM-pCAS-KO1-F (5'-fluo)	TGACGTCGCCGCCCATCAC
	pCAS-2R	ATTGGTTGTTGAGTTGGTTGTC

RT-PCR products	pSPL3-SD6	TCTGAGTCACCTGGACAACC
	6FAM-pSPL3-SD6 (5'-fluo)	TCTGAGTCACCTGGACAACC
	pSPL3-SA2	ATCTCAGTGGTATTTGTGAGC

Table S9. Primers used in the minigene splicing reporter assays.

Purpose	Gaussian distribution (Shapiro Wilk test)	Statistical analysis	Bioinformatics approach
Linear correlation between exon inclusion levels and *in silico* predictions	Yes	Spearman	QUEPASA, HEXplorer
	No	Pearson	SPANR, HAL, LR_{skip}, LR_{inc}
Discrimination of 3 groups of variants (↑ exon skipping *versus* no effect on splicing versus ↑ exon inclusion).	Yes	Anova (Bonferroni post-tests)	QUEPASA, HEXplorer
	No	Kruskal-Wallis (Duns post-tests)	SPANR, HAL, LR_{skip}, LR_{inc}
Discrimination of 2 groups of variants (variants that increase exon skipping or exon inclusion versus those that do not)	Yes	Student	QUEPASA, HEXplorer
	Yes	Student with Welsh's correction	
	No	Mann-Whitney	SPANR, HAL, LR_{skip}, LR_{inc}

Table S10. Statistical analyses conducted in this study.

Experimental subsets	New *in silico* approaches																				*In silico* approach combination															
	QUEPASA (-0.50)					HEXplorer (-14%)					SPANR (-0.1%)					HAL (-3.4%)					QUEPASA&HAL					AT LEAST 3					LR_{skip} (31.1%)				
	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc
ABCB11 n=82 (80)	75	63	**39**	89	66	65	58	**33**	84	60	55	57	**30**	79	56	**40**	79	**38**	80	70	40	84	**44**	81	73	50	73	**38**	81	67	70	67	**41**	88	68
BRCA1 Ex5* n=98	72	91	78	89	86	72	91	78	89	86	79	70	52	89	72	96	55	53	96	69	79	88	74	91	86	79	84	68	91	83	66	97	95	87	88
BRCA1 Ex6 n=43	100	67	**24**	100	70	100	62	**21**	100	65	100	51	**17**	100	56	100	62	**21**	100	65	100	74	**29**	100	77	100	77	**31**	100	79	100	69	**25**	100	72
BRCA1 Ex11 n=125 (121)	92	65	**22**	99	67	**17**	64	**5**	89	59	**42**	90	**31**	94	86	83	59	**17**	97	62	83	71	**23**	98	72	33	71	**11**	90	67	75	65	**19**	96	66
BRCA1 Ex18 n=28	100	50	**31**	100	59	100	75	**40**	100	79	75	67	**27**	94	68	75	92	60	96	89	75	92	60	96	89	75	79	**38**	95	79	100	75	**40**	100	79
BRCA1 n=14 (13)	80	78	67	88	79	80	56	50	83	64	80	63	57	83	69	80	78	67	88	79	80	89	80	89	86	80	78	67	88	79	80	75	67	86	77
BRCA2 Ex7 n=34 (29)	100	87	79	100	91	91	74	63	94	79	64	83	70	79	76	100	83	73	100	88	100	91	85	100	94	100	94	92	100	97	100	83	79	100	90
BRCA2 Ex18 n=24 (19)	100	**43**	50	100	64	88	50	50	88	64	**38**	91	75	67	68	100	50	59	100	71	100	50	59	100	71	90	64	64	90	75	100	**45**	57	100	68
BRCA2 n=51 (39)	71	71	55	83	71	65	59	**44**	77	61	75	62	**32**	91	64	63	79	59	82	74	47	91	73	78	76	59	62	**43**	75	61	75	66	**38**	91	68
CFTR Ex9 n=44 (41)	65	88	81	75	77	60	75	67	69	68	74	68	67	75	71	80	67	67	80	73	60	88	80	72	75	65	86	81	73	76	79	86	83	83	83
CFTR Ex12 n=41	68	89	88	71	78	86	84	96	84	85	82	63	72	75	73	95	84	88	94	90	68	89	88	71	78	82	89	90	81	85	82	89	90	81	85
DYSF *n=23*	**0**	65	**0**	81	57	**0**	70	**0**	82	61	**0**	50	**0**	77	**44**	**0**	100	**0**	87	87	**0**	100	**0**	87	87	**0**	85	**0**	85	74	**0**	80	**0**	84	70
F9 Ex5 n=17	83	100	100	71	88	83	100	100	71	88	92	60	84	75	82	83	80	91	67	82	75	100	100	63	82	75	100	100	63	82	75	80	90	57	76
FAS Ex6 n=171	59	87	73	78	77	59	91	79	79	79	56	68	50	72	63	82	81	70	90	82	54	94	83	77	79	57	93	81	78	79	63	89	77	81	80

MAPT Ex10* n=30	93	81	81	93	87	71	100	100	80	87	69	79	75	73	74	71	94	91	79	83	64	94	90	75	80	71	94	91	79	83	79	100	100	84	90
MLH1 Ex10 n=15	86	88	86	88	87	57	63	57	63	60	100	63	70	100	80	86	63	67	83	73	71	100	100	80	87	86	88	86	88	87	86	63	67	83	73
MLH1 n=71 (69)	89	63	**26**	98	66	78	65	**24**	95	66	75	61	**20**	95	63	67	85	**40**	95	83	67	85	**40**	95	83	89	92	**67**	98	82	88	69	**27**	98	71
MSH2 Ex5* n=22	92	100	100	91	95	92	90	92	90	91	83	80	83	80	82	92	100	100	91	95	92	100	100	91	95	92	100	100	91	95	92	100	100	91	95
MSH2 n=31 (30)	100	78	**40**	100	81	100	70	**33**	100	74	75	77	**33**	95	77	50	85	**33**	92	81	50	93	50	96	90	80	82	**38**	96	81	75	77	**33**	95	77
NF1 Ex9 n=36	63	94	92	70	78	79	76	79	76	78	89	41	63	78	67	68	76	76	68	72	58	94	92	67	75	74	88	88	75	81	79	88	88	79	83
NF1 Ex37 n=24	67	94	80	89	88	83	67	**45**	92	71	67	61	**36**	84	63	83	67	**45**	92	71	67	100	100	90	92	67	83	57	88	79	67	94	80	89	88
SMN1 Ex7 n=36	88	96	88	96	94	88	74	**47**	95	77	88	79	54	96	81	75	79	50	92	78	75	96	86	93	92	88	89	70	96	89	88	86	64	96	86
SMN2 Ex7 n=43	75	100	100	98	98	50	92	50	97	91	75	90	43	97	88	100	92	57	100	93	75	100	100	98	98	75	97	75	97	95	75	97	75	97	95
WT1 Ex5 n=139	75	86	71	88	83	61	71	**49**	80	68	77	68	53	87	71	75	57	**45**	83	63	64	92	78	84	83	64	81	61	83	76	70	84	67	86	79
Others n=122	85	**48**	74	66	72	83	59	78	67	75	72	61	77	55	68	66	77	84	57	70	65	77	84	56	70	74	73	83	62	74	85	55	77	67	74

Table S11. Comparative analysis of the performance of SRE-dedicated bioinformatics approaches in predicting variant-induced exon skipping in each of subset of the training and validation datasets.

Experimental datasets	New *in silico* approaches																				*In silico* approach combinations																	
	QUEPASA (+0.36)					HEXplorer (+9)					SPANR (+0.3%)					HAL (+1.0%)					QUEPASA&HAL					AT LEAST 3					LR_{inc} (6.2%)							
	Sen	Spe	PPV	NPV	Acc	Sen	Spe	Sen	Spe	PPV	NPV	Acc	Sen	Spe	Sen	Spe	PPV	NPV	Acc	Sen	Spe	Sen	Spe	PPV	NPV	Acc	Sen	Spe	Sen	Spe	PPV	NPV	Acc	Sen	Spe	PPV	NPV	Acc	Sen	Spe
BRCA1 Ex5* n=98	88	75	64	92	80	79	82	68	88	81	81	91	84	89	87	87	68	58	92	74	85	54	48	88	64	88	89	81	94	89	82	86	**75**	90	85					
BRCA1 Ex6 n=43	50	77	**33**	87	72	50	74	**31**	87	70	**0**	86	**0**	79	70	88	71	**41**	96	77	50	89	50	89	81	**25**	89	**33**	84	77	75	77	**43**	93	77					
BRCA1 Ex11 n=125 (121)	67	64	**4**	99	64	100	61	**6**	100	62	**0**	97	**0**	97	95	100	52	**5**	100	54	67	68	**5**	99	68	67	74	**6**	99	74	100	58	**6**	100	59					
CFTR Ex9 n=44 (41)	57	77	53	79	70	57	73	50	79	68	75	83	64	89	80	71	73	56	85	73	50	83	58	78	73	57	83	62	81	75	71	80	63	86	77					
FAS Ex6 n=171	75	63	**38**	89	66	90	57	**39**	95	65	75	**46**	30	86	53	85	56	**37**	92	63	70	69	**41**	88	70	80	64	**41**	91	68	78	59	36	90	63					
MAPT Ex10* n=30	67	87	83	72	77	93	80	82	92	87	77	71	71	77	74	93	80	82	92	87	67	93	91	74	80	80	83	92	85	87	73	93	92	78	83					
MLH1 Ex10 n=15	67	83	50	91	80	67	92	67	92	87	67	83	50	91	80	67	75	**40**	90	73	67	100	100	92	93	67	100	100	92	93	67	92	67	92	87					
NF1 Ex9 n=35	71	62	**31**	90	64	29	79	**25**	82	69	43	79	**33**	85	72	71	69	36	91	69	71	79	**45**	92	78	57	86	50	89	81	71	69	36	91	69					
SMN2 Ex7 n=43	77	69	85	56	74	80	62	83	57	74	90	**31**	75	57	72	93	54	82	78	81	100	77	91	100	93	83	62	83	62	77	97	62	85	89	86					
Total n=1217 (1182)	67	71	53	82	70	47	66	**40**	72	60	71	54	**43**	80	60	**47**	63	**38**	71	58	**33**	83	**48**	72	67	**42**	72	**42**	72	63	**42**	69	**42**	69	60					

Table S12. Comparative analysis of the performance of SRE-dedicated bioinformatics approaches in predicting variant-increased exon inclusion in subsets of the training and validation datasets.

Experimental subsets	New *in silico* approaches																*In silico* approach combinations									
	QUEPASA				HEXplorer				SPANR				HAL				QUEPASA&HAL			AT LEAST 3			LR_{skip}			
	Thr	Sen	Spe	Acc	Thr	Sen	Spe	Acc	Thr	Sen	Spe	Acc	Thr	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Thr	Sen	Spe	Acc
ABCB11 n=82 (80)	-0.63	70	**70**	**70**	-24.9	65	**73**	**71**	-0.11	55	57	56	-1.10	**60**	61	61	45	76	68	**55**	**84**	**77**	31.8	70	**69**	**70**
BRCA1 Ex5* n=98	-0.19	**90**	87	**88**	-16.8	**86**	75	79	-1.0	**83**	69	**74**	-1.5	76	**75**	**76**	76	**93**	**88**	**83**	**88**	**87**	24.2	**86**	80	82
BRCA1 Ex6 n=43	-1.9	75	**97**	**95**	-57.9	100	**95**	**95**	-1.19	75	**97**	**95**	-32.5	100	**82**	**84**	75	**97**	**95**	100	**97**	**98**	66.4	100	**92**	**93**
BRCA1 Ex11 n=125 (121)	-1.05	75	**74**	**74**	3.9	**50**	48	48	-0.04	**50**	70	68	-22.5	67	**79**	78	58	**84**	**82**	**42**	**76**	**73**	31.0	75	**67**	**68**
BRCA1 Ex18 n=28	-0.81	100	**79**	**82**	-47.6	100	**96**	**96**	-1.5	75	**83**	**82**	-1.8	**100**	88	89	**100**	75	79	**100**	**96**	**96**	38.3	100	**88**	**89**
BRCA1 n=14 (13)	-0.50	80	78	79	-23.2	80	**78**	**79**	-0.56	80	**78**	**79**	-6.2	80	**89**	**86**	80	**100**	**93**	80	**100**	**93**	28.9	80	**78**	**79**
BRCA2 Ex7 n=34 (29)	-0.65	100	**91**	**94**	-28.1	91	**83**	**85**	0.35	**73**	78	76	-5.0	91	**91**	**91**	91	**94**	93	91	94	93	53.7	91	**100**	**97**
BRCA2 Ex18 n=24 (19)	-1.38	80	**79**	**79**	-42.2	80	**71**	**75**	-0.06	**63**	91	**79**	-46.2	80	**79**	**79**	80	**79**	**79**	80	**79**	**79**	63.2	80	79	79
BRCA2 n=51 (39)	-0.45	71	71	71	-21.3	59	**65**	**63**	-0.23	75	**74**	**74**	-3.2	**71**	79	**76**	**53**	91	**78**	29	**82**	**65**	31.1	**76**	**68**	**71**
CFTR Ex9 n=44 (41)	-0.04	**80**	79	**80**	-0.80	**75**	67	**70**	-0.20	74	**73**	**73**	-12.1	75	**82**	**79**	**70**	88	**80**	**75**	**96**	**86**	22.9	**80**	83	82
CFTR Ex12 n=41	-0.24	**91**	84	**88**	-12.2	86	84	85	-0.83	74	**79**	**76**	-13.3	91	**89**	90	**86**	89	**88**	**86**	**95**	**90**	23.2	**91**	89	**90**
DYSF *n=23*	0.23	**67**	45	48	8.6	**67**	45	48	0.31	**33**	50	**48**	-1.1	**67**	68	65	**67**	70	70	**67**	65	65	18.3	**67**	45	48
F9 Ex5 n=17	-0.65	83	100	88	-8.3	**92**	100	**94**	-1.86	67	**80**	71	6.4	**100**	80	**94**	**83**	80	82	**83**	100	**88**	18.9	**100**	80	**94**
FAS Ex6 n=171	-0.06	**76**	76	76	3.8	**81**	80	**80**	1.9	**65**	63	64	-2.05	**76**	79	78	**67**	90	**81**	**76**	83	**81**	21.8	**79**	81	80
MAPT Ex10* n=30	-0.63	93	**94**	**93**	20.3	**93**	94	**93**	0.9	**85**	71	**78**	5.2	**100**	88	**93**	**93**	94	**93**	**100**	91	**97**	22.0	**93**	94	**93**

MLH1 Ex10 n=15	**-0.46**	**86**	**88**	**87**	**-24.7**	**57**	**75**	**67**	**-0.35**	**86**	**88**	**87**	**-4.8**	**86**	**63**	**73**	**71**	**100**	**87**	**86**	**100**	**93**	**20.0**	**100**	**63**	**80**
MLH1 n=71 (69)	-0.85	89	**73**	**75**	-50	**78**	**92**	**90**	-11	50	**100**	**94**	-8.2	44	**90**	**85**	44	**92**	**86**	57	**92**	**87**	50.9	78	**92**	**90**
MSH2 Ex5* n=22	0.6	**100**	90	95	-18.9	**100**	90	**95**	-0.15	83	**90**	**86**	-1.5	92	100	95	92	100	95	**100**	100	**100**	38.0	92	100	95
MSH2 n=31 (30)	-0.7	75	**81**	81	-42.3	75	**93**	**90**	-0.67	75	**92**	**90**	-15.4	50	**96**	**90**	50	**96**	**90**	75	**100**	**97**	28.2	**100**	**78**	**81**
NF1 Ex9 n=36	0.04	**68**	71	69	-16.8	79	**94**	**86**	-1.6	79	**88**	**83**	-10.5	68	**82**	**75**	63	88	75	74	**100**	86	34.9	76	**94**	85
NF1 Ex37 n=24	0.41	**83**	72	75	-30	83	**100**	**96**	-0.43	67	**72**	**71**	-20.5	83	**89**	**88**	**83**	89	88	**83**	**94**	**92**	28.9	**83**	89	88
SMN1 Ex7 n=36	-0.73	88	96	94	-28.6	75	**82**	**81**	-0.04	88	79	81	-3.9	75	79	78	75	96	92	88	89	89	31.0	88	86	86
SMN2 Ex7 n=43	-0.5	75	100	98	-6	**100**	90	91	0.19	**100**	90	**91**	-2.7	100	92	93	75	**100**	**98**	**100**	**97**	**98**	18	**100**	**95**	**95**
WT1 Ex5 n=139	-0.29	**84**	81	82	-2.0	**75**	59	64	-1.55	77	**76**	**76**	-8.2	70	67	68	61	89	81	**73**	79	**77**	24.5	**84**	80	**81**
Others n=122	-1.25	**72**	**70**	71	-21.7	78	**73**	**76**	-0.2	67	**64**	66	-2.0	**74**	59	69	62	**80**	68	**77**	**84**	**80**	43.0	74	**77**	**75**

Table S13. Optimisation of the thresholds of SRE-dedicated approaches for predicting variant-induced exon skipping in each subset of the training and validation datasets.

Experimental subsets	New *in silico* approaches																*In silico* approach combinations									
	QUEPASA (+0.36)				HEXplorer				SPANR				HAL				QUEPASA&HAL			AT LEAST 3			LR_{inc}			
	Thr	Sen	Spe	Acc	Thr	Sen	Spe	Acc	Thr	Sen	Spe	Acc	Thr	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Thr	Sen	Spe	Acc
BRCA1 Ex5* n=98	0.5	85	**81**	**83**	4.7	79	82	81	0.5	81	**93**	**88**	6	**76**	**77**	**77**	67	91	83	82	**91**	88	6	**85**	**88**	**87**
BRCA1 Ex6 n=43	0.23	**75**	74	**74**	0.85	**75**	69	70	0.05	**38**	60	56	5.8	75	**77**	77	**63**	**89**	**84**	**63**	69	67	7.1	75	77	77
BRCA1 Ex11 n=125 (121)	1.0	**67**	**74**	**74**	32.2	100	**74**	**74**	0.02	**67**	69	69	24.8	**100**	**69**	**70**	**67**	**81**	**81**	**100**	**83**	**83**	2.0	67	**81**	**81**
CFTR Ex9 n=44 (41)	0.22	73	**71**	**72**	4.6	64	**70**	68	0.32	75	86	83	0.65	71	73	73	**79**	83	**82**	**71**	76	75	7.1	**79**	80	**80**
FAS Ex6 n=171	0.57	73	**71**	**73**	27.5	70	**70**	**70**	3.4	60	**60**	**60**	20	73	**73**	**73**	60	**82**	**77**	60	**79**	**75**	4.4	70	**71**	**71**
MAPT Ex10* n=30	-0.63	**93**	**87**	**90**	20.3	**87**	87	87	0.90	69	**86**	**78**	5.1	87	**87**	87	**87**	93	**90**	80	**93**	87	9	**87**	87	**87**
MLH1 Ex10 n=15	0.85	67	**92**	**87**	44.8	67	**100**	**93**	-0.1	67	**92**	**87**	-4.8	**100**	75	**80**	67	100	93	67	100	93	9.7	**100**	83	87
NF1 Ex9 n=35	0.57	71	**72**	**72**	-10	71	**76**	**75**	-0.7	**86**	72	**75**	-3.2	**71**	**79**	**78**	71	**93**	**89**	**71**	83	81	5.5	71	**72**	**72**
SMN2 Ex7 n=43	0.64	70	**85**	74	14.5	80	**69**	**77**	5.1	70	**62**	67	12.5	73	**77**	74	53	**92**	60	73	**92**	**79**	3.2	70	**69**	70
WT1 Ex5 n=139	0.27	69	**70**	70	-3.1	**56**	55	55	1.8	60	**61**	60	-3.9	**56**	55	55	**38**	80	66	**49**	**82**	**71**	8.3	**62**	62	**62**

Table S14. Optimisation of the thresholds of SRE-dedicated bioinformatics approaches in predicting variant-increased exon inclusion in subsets of the training and validation datasets.

CHAPITRE II : ETUDES MUTATIONNELLES EXHAUSTIVES SUR DEUX EXONS MODELES DE CONCORDANCE-DISCORDANCE

Nos travaux portant sur l'évaluation à large échelle des approches de prédictions des altérations des ESR suggèrent qu'il est possible de prédire ce type d'altérations par des nouvelles méthodes bioinformatiques. Cependant, des résultats additionnels obtenus dans d'autres exons et/ou sur d'autres gènes ont également rapporté plusieurs cas discordants (données non publiées). L'ensemble de ces données laisse suggérer que ces méthodes de prédictions pourraient ne pas s'appliquer de manière équivalente à tous les exons et/ou à tous les gènes. Nous avons fait l'hypothèse que certains exons possèdent des caractéristiques spécifiques qui leur confèrent une sensibilité ou une résistance particulières aux mutations ESR permettant d'expliquer la concordance/discordance des résultats expérimentaux avec les prédictions bioinformatiques axées sur les ESR.

Pour tester cette hypothèse, nous avons utilisé comme modèles d'étude (i) l'exon 7 de *BRCA2* pour lequel une étude préliminaire a mis en évidence une fraction importante de mutations exoniques altérant des ESR potentiels (11/36, 31%) et (ii) l'exon 7 de *MLH1*, pour lequel aucune mutation de régulation d'épissage n'a été jusqu'ici identifiée. La stratégie a été d'examiner expérimentalement tous les variations exoniques identifiées dans ces exons pour leur l'effet sur l'épissage puis de confronter ces données avec celles obtenues *in silico* par les approches de prédictions d'anomalies d'épissage axées sur les ESR. Au total, 104 et 32 variations exoniques répertoriés dans les exons 7 de *BRCA2* et *MLH1*, respectivement, ont été extraites de l'ensemble des bases de données publiques puis ont été analysées à l'aide de tests fonctionnels d'épissage basés sur l'utilisation de minigènes et l'étude du matériel biologique du patient.

Dans un premier temps, nos résultats obtenus sur l'exon 7 de *BRCA2* ont permis d'identifier un nombre surprenant de mutations (80/104, 77%) affectant l'épissage de cet exon, en particulier via l'altération des ESR, suggérant que cet exon était particulièrement sensibles aux mutations touchant les ESR. En effet, la cartographie fonctionnelle des régions régulatrice d'épissage au sein de cet exon par une stratégie de marche sur exon a montré que celui-ci était particulièrement riche

en éléments régulateurs. Nos données ont également montré que l'effet de la plupart des variations sur l'épissage pouvait être prédit par les outils de prédictions axés sur les ESRs, validant le pouvoir prédictif des stratégies de filtration focalisée sur l'épissage. La relevance biologique des défauts d'épissage observés et des prédictions a d'ailleurs pu être confirmée *in vivo* lorsque l'ARN des patients était disponible. De plus, des analyses complémentaires, basées notamment sur la collecte des données génétiques, cliniques, tumorales et familiales des patientes porteuses des variations étudiées sont en cours afin d'évaluer le caractère pathogène des variations conduisant à un saut de l'exon 7 et de statuer sur leur pathogénicité. Ces données viennent compléter d'autres analyses au niveau protéique et ont des implications dans l'interprétation des variations.

Dans un second temps, nos résultats obtenus sur l'exon 7 de *MLH1* ont montré qu'aucune des mutations exoniques analysées dans cet exon n'altéraient l'épissage de l'exon 7 de *MLH1*, et malgré des prédictions en faveur d'une altération des éléments exoniques régulateurs de l'épissage. Ces données ont révélé que cet exon est particulièrement résistant aux mutations altérant des ESR et ont mis en évidence l'échec apparent des approches de prédictions axés sur les ESR. De plus, nos données ont contribué à expliquer la résistance de cet exon aux mutations affectant les ESR en démontrant l'implication des sites d'épissage, et notamment de leur force particulièrement élevée, qui explique également les discordances observées entre les prédictions bioinformatiques et les résultats expérimentaux. D'ailleurs, le pouvoir prédictif des approchées de prédictions dédiées aux ESR a finalement été restauré en diminuant la forces des sites d'épissage.

L'ensemble de ces données devraient contribuer à l'amélioration des outils bioinformatiques axés sur les ESR, approches permettant l'identification des mutations d'épissage potentiellement pathogènes parmi le très grand nombre de VSI actuellement détectées en diagnostic moléculaire.

L'ensemble des travaux menés sur l'exon 7 de *MLH1* font l'objet d'une publication qui sera prochainement soumise à *PLOS Genetics* (IF 6.100), tandis que les travaux portant sur l'exon 7 de BRCA2 sont encore en cours.

A staggering number of genetic variations affect the splicing pattern of *BRCA2* exon 7: validation of the predictive power of splicing-dedicated silico analyses

Tubeuf H[1,2*†], Caputo S.M[3*†], Abuli A[1], Moncoutier V[3*], Léoné M[4*], Boutry Kryza N[4*], Krieger S[1,5*†], Privat M[6*], Urhmmaer N[6*], Radice P[7†], Blok MJ[8†], Muller D[9*], Carré-Pigeon F[10*], Lazaro C[11†], Menéndez M[11], Wappenschmidt B[12†], Baralle D[13†], Eccles D[13†], Thomas S[14], Vega A[15†], Evans DG[16], Solano A[17†], Caligo M[18†], Drouet A[1], Karam R[18†], Pesaran T[18†], Hogervorst FB[19], Wijnen J[20], Wreeswijk M.P.G[21†], Revillon F[22*], Delnatte C[23*], Guillaud-Bataille M[24*], Bronner M[25*], Toland A[30†], Pedersen IS[27], Mitchell G[28,29], Driessen R[28], Shanley S[29], Hansen TV[30†], Peixoto A[31†], Teixeira M[32†], Claes K[33†], Frebourg T[1, 34], Gaildrat P[1*†], Spurdle A[35†], Martins A[1*†], on behalf of the ENIGMA consortium

[*] Unicancer Genetic Group (UGG) splice network
[†] ENIGMA consortium

1. Inserm-U1245, UNIROUEN, Normandie University, Normandy Centre for Genomic and Personalized Medicine, Rouen, France,
2. Interactive Biosoftware, Rouen, France
3. Institut Curie, Département de Biopathologie, Paris, France.
4. Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon-Centre Léon Bérard, Lyon, France.
5. Laboratoire de Biologie et Génétique du Cancer - Centre Normand de Génomique Médicale et Médecine Personnalisée, Centre François Baclesse, Caen, France.
6. Department of Oncogenetics, Jean Perrin Comprehensive Cancer Center, Clermont-Ferrand, France
7. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale Tumori, Milan, Italy.
8. Department of Clinical Genetics: GROW-School for Oncology and Developmental Biology, Maastricht University Medical Centre, 6229HX Maastricht, The Netherlands
9. Laboratoire d'oncogénétique, Centre Paul Strauss, Strasbourg, France.
10. Service de Génétique HMB, CHRU Reims, Reims, France.
11. Hereditary Cancer Program, Joint Program on Hereditary Cancer, Catalan Institute of Oncology, IDIBELL campus in Hospitalet de Llobregat, Catalonia, Spain.
12. Center for Hereditary Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), Medical Faculty, University Hosptial Cologne, Cologne 50931, Germany.
13. Human Development and Health, Faculty of Medicine, University of Southampton, Southampton S016 5YA, UK.
14. Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, UK
15. Fundacion Publica Galega de Medicina Xenómica-SERGAS Grupo de Medicina Xenómica-USC, IDIS, CIBERER, Santiago de Compostela 15706, Spain.

16. Genomic Medicine, Manchester Academic Health Sciences Centre, University of Manchester, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK.
17. INBIOMED, Faculty of Medicine, University of Buenos Aires/CONICET and CEMIC, Department of Clinical Chemistry, Medical Direction, Buenos Aires, Argentina
18. Section of Genetic Oncology, Department of Laboratory Medicine, University and University Hospital of Pisa, Pisa, Italy.
19. Ambry Genetics, Aliso Viejo, CA 92656, USA.
20. Netherlands Cancer Institute, Amsterdam, the Netherlands
21. Dept. Human Genetics, Leiden University Medical Center, the Netherlands.
22. Department of Clinical Genetics, Leiden University Medical Centre, Leiden 2300, The Netherlands.
23. Centre Oscar Lambret, Unité d'Oncologie Moléculaire Humaine, Lille, France.
24. Institut de Biologie, Laboratoire de Génétique Moléculaire, Service de Génétique Médicale, CHU Nantes, Nantes, France
25. Département de Biopathologie, Service de Génétique, Gustave Roussy, Université Paris-Saclay, Villejuif, F-94805, France.
27. Division of Human Cancer Genetics, Departments of Internal Medicine and Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, Ohio State University, Columbus, Ohio, USA.
28. Section of Molecular Diagnostics, Clinical Biochemistry, Aalborg University Hospital, Aalborg, Denmark.
29. Familial Cancer Centre, Peter MacCallum Cancer Centre, Department of Oncology, University of Melbourne, Australia.
30. Sir Peter MacCallum Dept. of Oncology, University of Melbourne, Parkville, VIC, Australia.
31. Department of Genetics, Portuguese Oncology Institute, Porto, Portugal.
32. Cancer Genetics Group, IPO Porto Research Center, Portuguese Oncology Institute of Porto, Department of Genetics, Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal.
33. Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium.
34. Department of Genetics, Rouen University Hospital, Normandy Centre for Genomic and Personalized Medicine, Rouen, France
35. Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

Author contributions

AM conceived and designed the project. HT and AA generated the experimental, bioinformatics and statistical data. SC performed familial data analysis. HT and AM were involved in data interpretation. HT and AM wrote the manuscript. All participants in this study are university-associated members of the Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) consortium, contributed with biological material and patient's data. All authors read and approved the final manuscript.

Fundings

This project was supported by the OpenHealth Institute, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (Gefluc) as well as the European Union and Région Normandie. Europe gets involved in Normandie with European Regional Development Fund (ERDF). HT was funded by a CIFRE PhD fellowship (#2015/0335) from the French Association Nationale de la Recherche et de la Technologie in the context of public-private partnership between INSERM and Interactive Biosoftware.

Abstract

Variant interpretation is a key issue in molecular diagnosis as identification of a causal mutation is essential for molecular diagnosis and clinical management of many genetic disorders. Spliceogenic variants exemplify this issue as each nucleotide variant can be deleterious by affecting RNA splicing signals (either splice sites or SREs, i.e. splicing regulatory elements). By testing an initial dataset of 36 variants in exon 7 of *BRCA2*, we have recently shown that a large fraction of VUS induce splicing defects, suggesting that *BRCA2* exon 7 was very sensitive to SRE-mutations which can predicted by newly developed splicing-dedicated *in silico* tools. Here, we extended our initial study by analyzing a larger number of new variants mapping to *BRCA2* exon 7 (n=66) or located in flanking intronic positions (n=15). Again, we resorted both to bioinformatics predictions and experimental analysis to evaluate the impact of each variant in RNA splicing. Our study revealed a staggering number of splicing mutations in *BRCA2* exon 7 (74% of the 121 analyzed variants), including mutations directly affecting splice sites and, particularly, mutations altering splicing regulatory elements. Importantly, our findings confirm the power of splicing-dedicated *in silico* analyses for prioritizing disease-causing candidates in the deluge of sequencing data. These data can complement further analyses at the protein level and have implications in variant interpretation.

Introduction

Today with the implementation of high-throughput DNA sequencing, one of the major challenges in medical genetics, notably in oncogenetics, is the interpretation of variants of unknown biological and clinical significance (VUS) (Cooper and Shendure, 2011; Frebourg, 2014). This is a key problem, especially in exome analyses, where thousands of nucleotide variations are typically identified per individual, as well as in the context of targeted sequencing of genes with large mutational spectra. The *BRCA* genes, implicated in familial predisposition to breast and ovarian cancer (HBOC, hereditary breast and ovarian cancer), embody the problem of variant interpretation. To date, most HBOC mutations have been identified in *BRCA1* (MIM #113705) and *BRCA2* (MIM #600185). Still, many suspected cases of HBOC remain without molecular explanation either because no sequence changes are observed in these genes or because VUS are detected (Eccles *et al.*, 2015). Depending on the database, VUS can represent over 30% of the variations detected in the *BRCA* genes in the Breast Cancer Information Core (BIC), ClinVar and BRCA Share databases (Caputo *et al.*, 2012; Landrum *et al.*, 2014; Szabo *et al.*, 2000). This is a critical issue for molecular diagnosis and appropriate clinical management of HBOC patients and their relatives, notably with the recent implementation of PARPi in the treatment of BRCA-deficient tumors (Ledermann *et al.*, 2014; Sonnenblick *et al.*, 2015).

The clinical relevance and the cancer associated risk of *BRCA1* and *BRCA2* VUS identified by genetic testing is frequently evaluated by using a multifactorial probability-based model that takes several parameters such as personal and family history of cancer, segregation of variants with cancer within families, co-occurrence with a known pathogenic variant, population-based case-control analysis and tumor histopathology features (Lindor *et al.*, 2012; Plon *et al.*, 2008; Tavtigian *et al.*, 2008). Unfortunately, there is often insufficient information to reclassify a VUS either as benign or pathogenic. Because most *BRCA* VUS are extremely rare, sometimes identified in small-size families for which segregation studies are difficult to perform outside particular research programs and access to clinical and family-based genetic information are frequently limited, functional data have recently been integrated for establishing the pathogenicity of genetic variants in inherited disease (Richards *et al.*, 2015). Numerous functional assays has been developed to evaluate the functional impact of *BRCA* variants and potentially infer their pathogenicity (Guidugli *et al.*, 2014). In parallel, the biological relevance of variants can be assessed using computational approaches. As VUS are typified by rare missense SNVs, prioritization strategies generally used

to pinpoint variants susceptible of causing disease rely mostly on *in silico* tools which focus on protein features (evolutionary conservation and biochemical properties of missense variants) (Adzhubei *et al.*, 2010; Kircher *et al.*, 2014; Ng and Henikoff, 2003; Tavtigian *et al.*, 2006). To exemplify, *in silico* predictions of functional impact of missense variants generated by A-GVGD are already incorporated in the multifactorial model proposed by the international consortium ENIGMA (Evidence-Based Network for the Interpretation of Germline Mutant Alleles), to classify missense variants in *BRCA* genes (Spurdle *et al.*, 2008; Tavtigian *et al.*, 2006). VUS also include variants potentially affecting mRNA splicing. Any intragenic VUS can be considered as potentially spliceogenic, including variations both within exons and introns either located at exon-intron junctions or outside of the canonical splice sites (Baralle *et al.*, 2009; Cartegni *et al.*, 2002; Wang and Cooper, 2007). It is now well established that a large fraction of *BRCA1/2* variants initially classified as VUS is in fact deleterious because they cause aberrant splicing by modifying splicing signals (Acedo *et al.*, 2012; Bonnet *et al.*, 2008; Houdayer *et al.*, 2012; Leman *et al.*, 2018; Sanz *et al.*, 2010; Théry *et al.*, 2011). Recently, a study based on a comprehensive analysis of all disease-causing splicing mutations reported in the Human Genome Mutation Database (HGMD) pointed to the existence of a group of 86 genes, especially affected by splice-site mutations (SSM) in which cancer genes such as *BRCA* genes were overrepresented. Yet, multifactorial model partially does not fully integrate the predicted effect of VUS at the mRNA level for the time being, especially those affecting SREs (Vallée *et al.*, 2016).

Over the past years, multiple *in silico* methods have been developed to predict the effect of VUS on splicing but have only been routinely applied for splice-site alterations (Houdayer *et al.*, 2012; Leman *et al.*, 2018; Moles-Fernández *et al.*, 2018). Nevertheless, numerous bioinformatics approaches aiming to predict potential variant-induced SRE (splicing regulatory elements) alterations have been recently described as promising tools to pinpoint spliceogenic exonic variants (Di Giacomo *et al.*, 2013; Erkelenz *et al.*, 2014; Ke *et al.*, 2011; Rosenberg *et al.*, 2015; Soukarieh *et al.*, 2016; Xiong *et al.*, 2015; Tubeuf *et al.*, in preparation). Especially, the predictive power of such *in silico* approaches have been demonstrated using an initial dataset of 36 variants within exon 7 of BRCA2, for which the effect on splicing was evaluated using a combined approach of minigene-based assays and patients' RNA analysis (Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012; Soukarieh *et al.*, 2016; Tubeuf *et al.*, in preparation). These previous studies thus suggested that

SRE-dedicated *in silico* tools could become useful in the stratification of exonic variants for functional analyses.

Here, we extended our initial study by analysing the entire set of nucleotide variations mapping to *BRCA2* exon 7 and additional variants located in flanking intronic positions. Again, we resorted both to bioinformatics predictions and experimental analysis to evaluate the impact of each variant on RNA splicing. This study validate the power of splicing-based *in silico* filtering strategies for prioritizing variants for functional analyses, this helping to identify potential pathogenic variants among the plethora of nucleotide changes detected by exome sequencing variants for functional analyses. These data can complement additional analyses at the protein level for a comprehensive assessment of pathogenicity and have implications in variant interpretation.

Material & Methods

Nomenclature. Nucleotide numbering is based on the cDNA sequence of BRCA2 (NM_000059.3), c.1 denoting the first nucleotide of the translation initiation codon, as recommended by the Human Genome Variation Society.

R**etrieval of *BRCA2* exon 7 variants.** A new collection of *BRCA2* variants identified within exon 7 and flanking intronic sequences was compiled within the ENIGMA consortium and also by interrogating 9 additional human databases (last accessed December 2017), namely: BIC (Breast Cancer Information Core, https://research/nhgri.nih.gov/bic/) (Szabo *et al.*, 2000), BRCA Share™ (Béroud *et al.*, 2016; Caputo *et al.*, 2012), ClinVar (https://www.ncbi.nlm. nih.gov/clinvar/) (Lek *et al.*, 2016), COSMIC (Catalogue of Somatic Mutations in Cancer, http://cancer.sanger.ac.uk/cosmic) (Forbes *et al.*, 2017), dbSNP (the Single Nucleotide Polymorphism database http://www.ncbi.nlm.nih.gov/SNP/), ESP (Exome Sequencing Project, http://evs.gs.washington.edu/EVS/), gnomAD (Exome Aggregation Consortium, http://exac. broadinstitute.org/), LOVD (Leiden Open Variation Database, http://chromium.lovd.nl/LOVD2/cancer/home.php?select_db=BRCA2) (Fokkema *et al.*, 2011) and HGMD (Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/index.php) (Stenson *et al.*, 2017) .

Splicing minigene reporter assays. In order to evaluate the impact of the selected *BRCA2* exon 7 variants on RNA splicing, we performed a functional assay based on the comparative analysis of the splicing pattern of wild-type (WT) and mutant reporter minigenes, as follows. Minigenes were prepared by using the pCAS2 vector (Gaildrat *et al.*, 2010; Soukarieh *et al.*, 2016) (Figure S1). Wild-type and mutant genomic fragments containing exon 7 and 159 bp and 208 bp of upstream and downstream introns, respectively (c. 516+58_c.631+208) were PCR amplified from patient genomic DNA by using forward B2Ex7_InFus_BglII-F and reverse B2Ex7_InFus_Mlu-R primers (Table S1) and then inserted into the BamHI and MluI cloning sites of the reporter plasmid pCAS2, yielding the three-exon hybrid minigenes pCAS2-*BRCA2*e7. When patient genomic DNA was not available, the variants of interest were introduced by site-directed mutagenesis by using the two-stage overlap extension PCR method (Ho *et al.*, 1989), a combination of specific primers indicated in Table S1 and the wild-type construct as template. Then, the mutant amplicons were introduced into pCAS2 by homologous recombination using the SLICE method (Motohashi, 2015). All constructs were sequenced to ensure that no unwanted mutations had been introduced into the inserted fragments during the PCR or cloning process. Next, WT and mutant minigenes (400 ng/well) were transfected in parallel into HeLa cells grown at ~70% confluence in 12-well plates using the FuGENE 6 transfection reagent (Roche Applied Science). HeLa cells obtained from ATCC were cultivated in Dulbecco's modified Eagle medium (Life Technologies) supplemented with 10% fetal calf serum in a 5% CO2 atmosphere at 37°C. Twenty-four hours later, total RNA was extracted using the NucleoSpin RNA II kit (Macherey Nagel) according to the manufacturer's instructions, and the minigenes' transcripts were analysed by fluorescent RT-PCR (30 cycles of amplification) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng total RNA and minigene specific primers (Table S1). RT-PCR products were separated by electrophoresis on 2.5% agarose gels containing ethidium bromide and visualized by exposure to ultraviolet light under saturating conditions using the Gel Doc XR image acquisition system (Bio-RAD), followed by gel-purification and sanger sequencing for proper identification of the minigene's transcripts. Finally, splicing events were quantitated by performing capillary electrophoresis on an automated 3500 Genetic Analyzer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) and computational analysis by using the GeneMapper v5.0 software (Applied Biosystems).

Branch point mapping. First, total RNA was extracted from a human lymphoblastoid cell line obtained from a healthy individual. Then, we incubated 1 µg DNase-treated RNA with RNase R (30U, Epicentre) for 30 min at 37°C. After purification, cDNA synthesis was performed with 500 ng RNase R-treated RNA using SuperScriptII™ Reverse Transcriptase (Life Technologies) and an outer reverse primer (Table S1). The first round of PCR (35 cycles of amplification) was set up using cDNA and FIREPol® DNA polymerase (SOLIS BIODYNE) with the outer forward and reverse primer set (Table S1). Then the mixture was divided into multiple reactions performed at different annealing temperatures (50-60°C). PCR products were pooled, purified and used as template for a second round of PCR (35 cycles of amplification) in which an inner primer set was used. Again, the mixture was divided and PCR reactions were performed at different annealing temperatures (50-60°C). PCR products were combined and run on an agarose gel. Bands of interest were excised, DNA extracted and purified and the products were ligated into the pGemTEasy cloning vector (Promega). Plasmid DNA from colonies were Sanger sequenced using a T7 Promoter primer.

Analysis of the *BRCA2* exon 7 splicing pattern in RNA samples from patients and controls individuals. The *BRCA2* exon 7 splicing pattern was analysed in RNA samples from patients and healthy control individuals isolated either from whole blood (LEU), primary cultures of stimulated peripheral blood lymphocytes (PBLs), short-term cultured peripheral blood lymphocytes (STCLs) or lymphoblastoid cell lines (LCLs) treated or not with puromycin, a NMD inhibitor (Table S2). The splicing pattern of *BRCA2* transcripts was analyzed by semi-quantitative fluorescent RT-PCR (see Table S2 for the number of amplification cycles) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng of total RNA and a combination of forward and reverse primers mapping to BRCA2 exons 6 and exon 9, respectively (Table S1). Then, RT-PCR products were separated by electrophoresis on a 2% agarose gel, gel-purified and sequenced and splicing events were quantitated by performing capillary electrophoresis on an automated sequencer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) and computational analysis by using the GeneMapper v5.0 software (Applied Biosystems).

Allele specific expression analysis. Allele specific expression (ASE) was measured by performing a SNaPshot quantitative primer extension assay (SNaPshot MultiplexKit, Applied Biosystem) (Soukarieh *et al.*, 2016). RT-PCR products spanning *BRCA2* exons 6 to 9 were obtained from

patients' RNA samples by using the forward RT-B2Ex6-F and reverse RT-B2Ex9-R primer set (Table S1). In parallel, the genomic segment encompassing *BRCA2* exon 7 was amplified by PCR from the genomic DNA of the same patient by using the forward BRCA2Ex7_InFus_Bam-F and reverse BRCA2Ex7_InFus_Mlu-R primer set (Table S1). The sequences of the variant-specific primers used in the primer extension reaction are indicated in Table S1. Purified extension products along with 120 LIZ Size Standard (Applied Biosystems) were separated by capillary electrophoresis on an automated 3500 Genetic Analyzer (Applied Biosystems) and analysed by using the GeneMapper v5.0 software (Applied Biosystems). SNaPshot results obtained from patient cDNA were normalized to those obtained from patient gDNA.

Splicing-dedicated bioinformatics predictions. Three types of bioinformatics methods were used to predict variant-induced splicing alterations, namely: splice site (ss)- , branch point (BP)- or splicing regulatory elements (SRE)- dedicated methods depending on the position of the variants relative to the exon.

For intronic variants and for those mapping to exon termini at positions overlapping the splice sites, we resorted to MaxEntScan (Yeo and Burge, 2004) (MES, http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html; MaximumEntropy Model), SpliceSiteFinder-like (Shapiro and Senapathy, 1987) (SSFL, Interactive Biosoftware, Rouen, France) and Splicing Prediction in Consensus Elements (SPiCE) (Leman *et al.*, 2018). For SPiCE, we used optimal sensitivity threshold (Thse) recommended for molecular diagnosis, i.e. variants were predicted to alter splicing if Thse, ≥ 11.5%. MES and SSFL were used both as stand-alone and in a combined sequential fashion (MES+SSFL) as recommended by Houdayer and colleagues (Houdayer *et al.*, 2012), i.e. variants were predicted to alter splicing if MES ≤ -15% and SFFL ≤ 5%. These algorithms were interrogated by using either the integrated software tool Alamut Batch v1.9 or Alamut Visual v2.11 (Interactive Biosoftware, http://www.interactive-biosoftware.com) whereas SPiCE scores were retrieved from the dedicated software tool (https://sourceforge.net/projects/spicev2-1/). The possibility of variant induced de novo splice sites/activation of cryptic splice sites was assessed by annotating all increments in local MaxEntScan and SpliceSiteFinder-like scores and comparing their values with those of reference splice sites. Only scores equal or higher to those of the corresponding reference splice site were considered as potentially spliceogenic.

The position of putative BPs, as well as the potential impact of variants mapping at these BPs was predicted by Branch pointer (Signal *et al.*, 2018), BPP (Zhang *et al.*, 2017), Human Splicing Finder (Desmet *et al.*, 2009) (HSF, http://www.umd.be/HSF/), SVM-BP finder (Corvelo *et al.*, 2010) and SROOGLE (http://sroogle.tau.ac.il/, Schwartz *et al.*, 2009) that provides branch site scores based on Kol *et al.* (Kol *et al.*, 2005) and on Schwartz *et al.* (Schwartz *et al.*, 2009).

For the prediction of variant-induced impact on ESR, we resorted to four newly developed SRE-dedicated *in silico* approaches: (i) the QUEPASA method previously described by Ke and co-workers (Ke *et al.*, 2011) and implemented by our group (Di Giacomo *et al.*, 2013), which is based in the calculation of total ESR_{seq} score changes ($\Delta tESR_{seq}$) (ii) the HEXplorer method (Erkelenz *et al.*, 2014) which calculates ΔHZ_{EI} values (iii) the SPANR approach described by Xiong and co-workers (Xiong *et al.*, 2015) which yields $\Delta\Psi$ scores, and (iv) HAL based on the calculation of $\Delta\Psi$ scores, as described by Rosenberg and co-workers (Rosenberg *et al.*, 2015)., Both $\Delta tESR_{seq}$ and ΔHZ_{EI} scores were calculated by using the Alamut Batch prototype tool version 1.5.2 (ESR_{seq}), (Interactive Biosoftware, Rouen, France), whereas SPANR and HAL scores were retrieved from the dedicated online interfaces (http://tools.genes. toronto.edu and http://splicing.cs.washington. edu/ SE, respectively). For each SRE-dedicated *in silico* tool, score changes (Δ) of the exonic variants, smaller than the indicated thresholds were considered predictive of increased exon skipping. Combinations of SRE-dedicated methods were performed as recently described (Tubeuf *et al.*, in preparation).

Statistical a**nalyses.** Results are presented as the mean ± SEM of three independent experiments. Data derived from confrontation of experimental and *in silico* analyses were compared by using either Student's test, one-way ANOVA test and Pearson's correlation coefficient or their derivatives depending on data distribution patterns as detailed in Table S3. In general terms, Mann-Whitney (non-Gaussian distribution) or Student's were used for assessing the performance of the bioinformatics tools in when only 2 groups of variants were taken into account (i.e. variants that increase exon skipping versus those that do not). Similarly, the Kruskal-Wallis or ANOVA tests followed by Duns or Bonferroni post-tests, respectively, were used for assessing the performance of the bioinformatics tools in discriminating 3 groups of variants (i.e. variants that increase exon skipping versus those with no effect on splicing versus those that increase exon inclusion). Linear Correlation between exon inclusion levels and *in silico* predictions was measured by calculating

Spearman or Pearson correlation coefficients (r). All statistical analysis were performed by using GraphPad Prism software (Version 5.0). Results are expressed as two sided p-values (* p-value<0.05, ** p-value<0.01, *** p-value<0.001) and were considered significant when p-value <0.05).

Performance assessment. The evaluation of the predictive power of splicing-dedicated bioinformatics methods was performed by measuring sensitivity (Sen) = [TPx100/(TP + FN)], specificity (Sp) = [TNx100/(TN + FP)], accuracy (Acc) = [(TN+TP)x100/(TN+TP+FN+FP), where TP (true positive) and FN (false negative) values are the numbers of positive samples that are predicted to be positive and negative respectively. Analogously, TN (true negative) and FP (false positive) values are the numbers of negative samples that are predicted to be negative and positive respectively. TP, TN, FP, FN were determined by taking into account thresholds determined either previously (Tubeuf *et al.*, in preparation), as indicated, or by performing new ROC (Receiver operating characteristic) curves representing the "closest to top left" (i.e. by minimizing the distance at the top left corner). ROC curves were performed by using GraphPad Prism software (Version 5.0) and easyROC (http://www.biosoft.hacettepe.edu.tr/easyROC/). The predictive power of the SRE-dedicated tools were then compared to each other by using Venn diagrams plotted by Jvenn (Bardou *et al.*, 2014), an interactive web application (http://jvenn.toulouse.inra.fr/app/ example.html).

Results

Selection of 81 variants newly reported within *BRCA2* exon 7 and its flanking intronic regions.

We have recently revealed that a high number of variants (15 out of 36, i.e. 42%) identified in the exon 7 of the *BRCA2* gene were shown to affect exon 7 splicing, mainly by altering potential ESR (11 out of 36, i.e. 31%), most of which could have been predicted by using new SRE-dedicated *in silico* tools (Di Giacomo *et al.*, 2013; Soukarieh *et al.*, 2016; Tubeuf *et al.*, in preparation). To further assess the prevalence of mutations that affect *BRCA2* exon 7 splicing and the ability of splicing-based *in silico* tools to predict such mutations, we extended our initial study by analysing the entire set of nucleotide variations mapping to *BRCA2* exon 7 and additional variants located in

flanking intronic positions. We began by collecting all variants located within *BRCA2* exon 7 and its flanking intronic regions reported by molecular diagnostic laboratories within the ENIGMA consortium (Table S4). In addition, we also included in our cohort all variants reported within *BRCA2* exon 7 by interrogating 9 human variation databases that list genetic changes identified either in the genomes of HBOC patients, in the general population or in tumors (Table S4). As a result, we collected a total of 81 new *BRCA2* variations for no splicing data had been obtained in minigene assays, including 63 single nucleotide variations (SNVs) and 18 small indels, most of which identified in cancer patients suspected of HBOC (Table S4). Only 19 of these variations are currently classified as unequivocally pathogenic and 6 as unequivocally not pathogenic, whereas the remaining 56 variations include either variations not yet classified (n= 23), VUS (n= 25) or variations with conflicting interpretation (n= 8) (Table S4). Altogether, this selection comprised 15 intronic variation, 3 of which mapping to the invariant splice site positions IVS ± 1/2 and 66 exonic SNVs most of which (n=64) being located outside the splice sites and herein considered as potential ESR alterations.

A large fraction of variants mapping to *BRCA2* exon 7 splice sites or to flanking intronic positions affect its splicing pattern.

We first focused our attention on the group of variants directly mapping to the splice sites of exon 7 or located on the flanking intronic positions. This subset represented a total of 17 variants, including 2 exonic and 15 intronic as shown in Figure 1 and Tables S4 and S5. In order to assess their impact on RNA splicing, we performed a cell-based splicing assay based on pCAS2-*BRCA2*e7-derived minigenes (Figure S1). As shown on Figure 1B and Tables S1, the wild-type (WT) pCAS2-*BRCA2*e7 minigene generated two different transcripts: one predominant transcript containing exon 7 (FL, 85% of exon inclusion) and a minor transcript without exon 7 (Δ7, 15% of exon skipping). These results are in agreement with those previously reported by using the same minigene system (Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012). Importantly, the minigene assay revealed that 13 out of the 17 variations (76%) altered the splicing pattern of exon 7 relative to wild-type by either increasing (n==11) or decreasing (n=2) exon 7 skipping (Figure 1 and Table S4). The 11 variants leading to an increase in Δ7 can be separated into two categories according to the severity of the splicing defect: (i) a first category consisting of 7 variants causing drastic splicing defects (≤5% FL), either near-total to total exon 7 skipping (n=6) or the deletion of the

first 1 nucleotide of the exon 7 concomitant to exon 7 skipping (Figure S2), most of which are located on the most conserved positions of the 3' ss (YAG│G, first exonic position underlined) and 5' ss (CAG│GURAGU, 3 last exonic positions underlined), and (ii) a second category consisting of 4 variants inducing moderate splicing defects (76% to 31% FL), possibly affecting branch site or intronic splicing regulators. Of note, 5 of the 7 variant causing severe splicing defects are currently classified as pathogenic, suggesting that the 2 remaining variants (c.517-13_-9delTCTT and c.631G>T) may be equally deleterious, whereas the 4 remaining variants causing partial defects should be considered as VUS (c.517-23_-22del, c.517-20A>G, c.517-19C>T and c.63+7A>G) (Table **S4).**

We then decided to compare these experimental data with splice site-dedicated *in silico* predictions in order to assess their accuracy in pinpointing variants that cause exon skipping. Because of the short sequence windows taken into account by MES+SSFL and SPiCE (i.e. 14 intronic and 2 exonic positions at the 3'ss and 3 exonic and 6 intronic positions at the 5'ss), only 9 out of the 17 variants could be analyzed by these methods (Table S5). We observed that 7/9 variants were correctly by the combination of MES+SSFL (78% accuracy) and 9/9 variants (100% accuracy) were correctly predicted by SPiCE (Figure 1C & Table S5). Besides inducing exon 7 skipping, some variants also caused deletions of exonic segments due to the creation of *de novo* 3' ss (c.517-1G>A) or activation of a cryptic 5'ss (c.631G>T, c.631+2T>C, c.631+3A>G) concomitant with the disruption of the corresponding splice site. These defects could not be anticipated by MES+SSFL nor SPiCE but can be explained when looking into local MES and SSFL scores, at least for c.517-1G>A (Figure S2). These cases highlight the importance of experimentally analyze predicted variant-induced splicing alterations to assess their actual outcomes and pinpoint the limits of SPiCE which is able to predict potential splicing alteration of variants at consensus splice sites but not the type of the effect (exon skipping or use of alternative/cryptic splice site).

We surmise that a large fraction of variants mapping to *BRCA2* exon 7 splice sites have an important impact on exon 7 splicing and that SPiCE outperformed MES+SSFL in predicting which of these variants induce splicing defects. Moreover, we found that *BRCA2* exon 7 appears to be particularly sensitive to splicing alterations as we found that additional variants located outside the splice sites in the upstream and downstream introns also have a negative effect on exon 7 inclusion. Given their position, these alterations could not be predicted by MES+SSFL nor by SPiCE

underscoring the importance of experimentally testing intronic variants flanking exon 7, and the need for complementary prediction tools dedicated to other splicing signals.

***BRCA2* exon 7 have several branch points contributing to its efficient splicing.**

We noticed from our minigene results that from the 8 variants tested in intron 6, 3 variations located upstream the acceptor splice site (c.517-23_-22del, c.517-20A>G and c.517-19C>T) increased exon skipping to different extends from 24 to 69%, depending on the variant) (Figure 2A). Given their position, we suspected that they could affect the branch point site of exon 7. We thus analyzed the sequence immediately upstream the 3'ss of *BRCA2* exon 7 by using five computational BP predictors (Branch pointer, BPP, HSF, SVM-BP finder and SROOGLE). As shown on Figure 2B, 6 putative BPs clustering near c.517-20 and c.517-50 positions were indicated by at least one BP predictor, namely c.517-21A, c.517-28A, c.517-29A, c.517-45A, c.517-48A and c.517-49A. Next, we decided to experimentally identify the exact position of the BP used for *BRCA2* exon 7 splicing by a mapping approach consisting in the amplification of the corresponding RNA lariat intermediate from a human lymphoblastoid cell line (Figure 2C). Sequencing of the lariat product obtained by nested RT-PCR (~ 150 bp) implicated three individual adenine, residues located at positions -29, -22 and -17, as real BPs (Figure 2D). Our results are consistent with several known features of human BPs, especially: (i) the position of the BP relative to the nearest 3'ss (within a window of ~40 nt upstream the 3'ss for 90% of human BPs), (ii) the type of nucleotide used as BP (an adenosine for 92% of human BPs), and (iii) introns can have multiple branchpoints generally clustered in close proximity to each other (Gao *et al.*, 2008; Mercer *et al.*, 2015; Pineda and Bradley, 2018). Interestingly, the c.517-22 adenine was recently identified by Mercer and co-workers (Mercer *et al.*, 2015). However, c.517-23_-22del did not totally abolished exon 7 inclusion in our minigene assay suggesting that the c.517-22 BP is important but not essential for *BRCA2* exon 7 splicing. We thus suspect that the 3 BP may play redundant roles in exon 7 splicing. We also noticed that none of the 3 detected branch sites conform to the yUnAy consensus BP motif, in particular they do not present the canonical uridine at position -2 relative to the BP. Yet, the introduction of an uridine by c.517-19C>T 2 nucleotides upstream of the BP adenines did not have a positive impact on exon 7 splicing, on contrary, it slightly increased exon skipping and thus seem to represent suboptimal splice sites. These data illustrate the difficulty of predicting variant-induced

alterations of BPs and highlight the importance of experimentally testing exonic variants mapping upstream for their impact on splicing.

BRCA2 exon 7 is highly sensitive to alterations of potential exonic splicing regulatory elements.

We then asked if the regulation of the splicing pattern of *BRCA2* exon 7 was affected by any of the new 64 variants located outside the splice sites reported in this exon, including 37 presumed missense, 4 nonsense, 8 synonymous and 15 frameshift variants (Table S4). given the good performances of the newly developed SRE-dedicated *in silico* approaches (QUEPASA, HEXplorer, SPANR and HAL) in predicting the impact on splicing of an initial dataset of *BRCA2* exon 7 variants (Di Giacomo *et al.*, 2013; Soukarieh *et al.*, 2016; Tubeuf et al, in preparation), we decided to first analyze the new 64 variants to by using these bioinformatics tools in order to identify those more susceptible to induce exon skipping by altering potential SRE. As shown in Table S6, and depending on the *in silico* tool taken into account, 14 to 27 out of these 64 variants (2-42%) were predicted to induce exon skipping by affecting SRE.

We then performed *ex vivo* splicing assays with pCAS2-*BRCA2*e7 derived minigenes to experimentally test all the variants and verify these predictions (Figure S1). As shown in Figure 3A and Table S4, that 23 out of the new 64 variants (36%) caused exon skipping, decreasing the level of full length transcripts to different extents, the strongest effect (≥ 95% exon skipping) being observed with c.523C>T (p.Gln175*), c.559G>T (p.Glu187*) and c.620C>G (p.Thr207Ser) and 23 out of the new 64 variants (36%) increased exon inclusion. Moreover, we identified 4 variations leading to the total (n=1, c.566A>G) or partial (n=3, c.583del, c.583_596delinsAGG and c.593_596delinsAGG) deletion of part of exon 7 at its 3' extremity (70 nt, 50 nt, 36 nt, and 22 nt, respectively), due to the usage of a de novo 5'ss in the case of c.566A>G or the use of a cryptic 5' ss (Figure S2), bringing to a total of 50 the number of variants that changes the splicing of exon 7 in this series (Table 1). Of note, only the splicing outcomes induced by c.566A>G would have been predicted by using both MES and SSFF (Figure S2). Thus, this variant was not considered as potential SRE-mutations and was no retained for downstream analysis. In contrast, the activation of the cryptic 5'ss in the context of c.583del, c.583_596delinsAGG and c.593_596delinsAGG was not predicted by MES and/or SSFL (Figure S2), underlying a potential alteration of SRE.

Altogether, our data revealed that a staggering number of variations affect the splicing pattern of *BRCA2* exon 7 and confirms that *BRCA2* exon 7 is particularly sensitive to variant-induced SRE alterations.

To get a better appreciation of the predictive power of the four SRE-dedicated *in silico* approaches (QUEPASA, HEXplorer, SPANR and HAL) and their different combinations (QUEPASA&HAL, AT LEAST 3 and LR_{skip}), in pinpointing variant-induced exon skipping, we then compared their outputs with the minigene results. This analysis was performed by calculating the number of positive/negative true/false calls produced by each methods based on the optimal thresholds previously determined for *BRCA2* exon 7 (i.e., -0.65 for QUEPASA, -28 for HEXplorer , +0.35% for SPANR , -5.0% for HAL and 33.6% for LR_{skip}, Table S6) (Tubeuf *et al.*, in preparation). As shown on Figure 3B and Table S6, QUEPASA (TC=54, 86%), HAL (TC=51, 81%) and HEXplorer (TC=48, 76%) produced the highest number of true calls (TC), outperforming SPANR (TC=30, 64%) in discriminating variants that induce exon skipping from those that do not. Overall, our results revealed better specificity, sensitivity and accuracy in predicting exon skipping for QUEPASA (74%; 93%; 86%), HAL (83%; 80%; 81%) and HEXplorer (47%; 93%; 76%) than for SPANR (69%; 61%; 64%) (Figure 3B & Table S6). Overall, we found that QUEPASA&HAL (TC=56, 89%) outperformed all the others approaches ($64\% \leq \text{Accuracy} \leq 86\%$) in discriminating variants that induce *BRCA2* exon 7 skipping from those that do not (Figure 3B & Table S6).

According to the functional splicing data obtained in the context of the pCAS2-*BRCA*2e7 minigene, the 63 exonic SNVs located outside splice sites into could be separated into three categories as follow: (i) variations that increased exon 7 skipping (n= 23), (ii) variations that do not affect the exon 7 splicing pattern (n= 15) and (iii) variations increasing exon 7 inclusion (n= 25) (Figure 3A; Tables S4 and S6). Statistical analyses showed that although all the approaches were able to discriminate variants that lead to exon skipping from those that to not (T-test, p-values <0.001), only QUEPASA and HEXplorer were able to predict the direction of the induced splicing defects, i.e. to discriminate the 3 categories of variants (Figure S3, ANOVA's post-test, p-value < 0.01 for QUEPASA and HExplorer and p-value "exon inclusion versus no effect" ns for SPANR, HAL and LR_{skip}). Nevertherless, we observed a statistically significant correlation between the level of exon inclusion detected in the minigene-based splicing assay and the score differences produced by

QUEPASA, HEXplorer, SPANR, HAL and LR_{skip} (Figure S4, Spearman correlation, p-values <0.001), indicating that each approach is able to estimate the severity of the splicing defects.

Optimal thresholds for SRE-dedicated predictions.

As suggested in our recently published large scale evaluation of SRE-dedicated *in silico* tools, the optimal threshold used to predict variant-induced exon skipping might vary depending on the intrinsic characteristics of the exon of interest, which may in turn influence our perception of the the reliability of such tools (Tubeuf *et al.*, in preparation). Therefore, we next wonder if we could improve the predictive power of SRE-dedicated *in silico* approaches by further optimizing the optimal decision thresholds (corresponding to the best compromise between specificity and sensitivity) relative to the full dataset of 97 *BRCA2* exon 7.

Surprisingly, none of the mutations tested in our previous study was responsible of increased exon inclusion (Di Giacomo *et al.*, 2013), probably due to a lack of sensitivity of the method used at the time to quantify the splicing defects and/or the eventual presence of heteroduplexes. In that study, we tested a total of 36 *BRCA2* exon 7 variants, some of which caused exon skipping (n = 15) and some do not (n = 21), as determined experimentally by using the same pCAS2 minigene-based assay (Di Giacomo *et al.*, 2013). To better assess the effect on splicing of this initial dataset, we decided to re-evaluate the splicing outcomes by using the same semi-quantitative fluorescent RT-PCR method already used for the new variants in this study. As described in Table1, here, we confirmed that 10 out of the 21 variants initially thought not to alter splicing (Di Giacomo *et al.*, 2013), here induced no effect on exon 7 inclusion; but that 11 variants show some changes relative to WT, albeit of low intensity. These include one variant that induces a weak increase in exon skipping (c.623T>G) and 10 variants that increase exon inclusion (Table S4). Hence, our results revealed a striking high proportion of splicing mutations within *BRCA2* exon 7 (80/104, 77%) similar to what was observed in *MLH1* exon 10 (17/22, 77%) (Soukarieh *et al.*, 2016).

By taking advantage of a ROC curve analyses (Figure S5), relative to the full datatset of 98 BRCA2 exon 7 variants, the best compromise between specificity and sensitivity was found at threshold of -0.32 for QUEPASA, -11 for HEXplorer , +0.38% for SPANR , -4.2% for HAL and 25.7% for LR_{skip} (Table S4), which were close to those determined from the large-scale comparative analysis,

i.e. -0.5, -14, -0.1%, -3.4% and 31.1%, respectively (Tubeuf *et al.*, in preparation). Accordingly, we observed a slight increase in accuracy (68-92% range versus 64-89% range), due to an increase in sensitivity (61%-98% range versus 67%-98%), while a good specificity is maintained (%-% range versus %-% range) when adjusting the threshold to the full set of *BRCA2* exon 7 variants (Table S4), suggesting that the decision threshold determined in Tubeuf *et al.*, might be applied prospectively to new series of BRCA2 exon 7 variants with good performances.

Our data confirm (i) the reliability of SRE-dedicated *in silico* approaches in pinpointing variant-induced ESR alterations and in predicting the severity of the induced splicing defects, and (ii) that the combination of such tools may improve the predictive power of ESR-dedicated bioinformatics analyses.

Splicing regulatory sequences within *BRCA2* exon 7 contribute to its efficient splicing.

To better understand how regulatory sequences are distributed within *BRCA2* exon 7, we performed a serial microdeletion analysis by using an exon-walking strategy in the context of the pCAS2-*BRCA2*e7 minigene assay (Figure S1). To this end, eleven *BRCA2* exon 7 segments lacking consecutive ~10 bp fragments (del1 to del11), sparing the first exonic and the three last nucleotides, were inserted in to the pCAS2 minigene in line of the *BRCA2* exon 7 wild-type sequence and their impact on exon definition was evaluated by analyzing the minigene transcripts expressed in transiently transfected cells (Figure 4A).

Our results revealed that all the 10-nt microdeletions alter the splicing pattern of *BRCA2* exon 7 when compared to pCAS2-*BRCA2*e7 WT, indicating that these 10-nt sequences probably contain regulatory elements guiding exon recognition (*Figure* 1B). More precisely, deletion del1 and del5 induced almost total exon 7 skipping ($\Delta 7 \geq 90\%$), suggesting that the corresponding 10 nt missing sequences provide a stronger contribution to the recognition of *BRCA2* exon 7 than those corresponding to del3, del6, del7, del9 and del11, possibly by being enriched in ESE or lacking ESS elements. In contrast, del2, del 4, del8 and del10 induce almost total exon 7 inclusion ($\Delta 7 \leq 7\%$) suggesting that the corresponding missing segments play a role in the recognition of *BRCA2* exon 7, probably by containing a higher density in ESE elements (Figure 1B). Interestingly, we observed that most of the *BRCA2* SVNs mapping in the exonic segment missing in del 5 induce exon skipping of exon 7 whereas those mapping to that of del8 induce an increase in exon inclusion,

broadly recapitulating the effects produced by the microdeletions. We hypothetized that the former destroy ESE elements whereas the latter disrupt ESS.

Confirmation of variant spliceogenic effect from RNA samples of HBOC patients.

To apprehend the physiological relevance of the splicing defects detected in the minigene assay and as well as the clinical value of splicing-dedicated in silio predictions, we compared our results with data derived from the analysis *of* RNA samples obtained from carriers of equivalent heterozygous BRCA2 variants (Table S3). We had the opportunity to collect 36 RNA samples from patients for 23 different SNVs, including 14 exonic variations and 9 intronic variations (Tables 1). In addition, we were able to obtain RNA samples from two related patients harboring a complete genomic deletion of BRCA2 exon 7 (c.517_c.631+462del) and used as control in our analysis (Tables 1).

RT-PCR characterisation of variant-induced BRCA2 exon 7 splicing defects in blood-derived samples. The analysis of blood-derived RNA samples was performed by semi-quantitative fluorescent RT-PCR by using primers targeting *BRCA2* exons 6 and 9. Several control samples, which were analysed in parallel of those of patients, revealed the production of a single transcript corresponding to normal exon 7 *inclusion* (FL, full length) (Tables 1). These data is consistent with previous descriptions of *BRCA2* splicing pattern in normal tissues (Davy *et al.*, 2017; Fackenthal *et al.*, 2016). In addition and as expected, analysis of RNA samples (LCL and PAXgene), from two individuals harbouring a heterozygous deletion of *BRCA2* exon 7 revealed, the presence of both normal transcripts containing exon 7 and aberrant transcripts lacking exon 7 (Δ7 = 34% and 58% in LCL and PAXgene, respectively) (Tables 1). Surprisingly, although LCL and PAXgene samples were collected from the same two individuals, the relative levels of Δ7 transcript levels were lower in LCL as compared to PAXgene samples (Table 1). Given that *BRCA2* exon 7 skipping produces out-of-frame Δ7 transcripts, we suspected that those aberrant transcripts may be degraded by the NMD pathway in LCLs. Accordingly, the relative level of Δ7 transcripts increased in LCLs in the presence of the NMD-inhibitor puromycin (Δ7 = 35% to 55% without and with puromycin, respectively), reaching a level similar to the one observed in the PAXgene samples (Δ7 = 58%), suggesting that NMD may be less effective in PAXgene-stabilized blood (Table 1).

The splicing pattern of *BRCA2* exon 7 of the biological samples obtained from individuals carrying the 23 SNVs of interest mapping to/near BRCA2 exon 7 were compared to those generated with equivalent RNA samples from at least 3 healthy control individuals. Results shown in Table 1 indicate that patients carrying either c.573T>C, c.575T>C, c.605C>G, c.625C>T, c.627C>T, c.631+25C>T, c.631+29A>C or c.631+43G>T had a splicing pattern similar to that of healthy controls, suggesting that these seven variations did not affect exon 7 splicing *in vivo*, in agreement with minigene data. In contrast, patients harbouring either c.517-13_-9del, c.517-2A>G, c.517G>T, c.520C>T, c.631G>A and c.631+3A>G variations exhibit a splicing pattern similar to that of patients carrying the heterozygous *BRCA2* exon 7 genomic deletion, i.e an apparent decrease in the amount of FL *BRCA2* transcripts and a concomitant drastic increase in exon 7 skipping (Δ7 = 45-65% and 55-58% for the variants of interest and the genomic deletion, respectively), as compared to healthy controls (Δ7 = 0.5%) (Table 1). Sequencing of the FL RT-PCR products of patients carrying the c.517G>T and c.631G>A exonic variations in parallel of the PCR products obtained with genomic DNA from the same patients revealed the absence of the mutant FL transcripts, suggesting that c.517G>T and c.631G>A cause near-total to total exon splicing defects (Figure S6). In contrast, sequencing of the FL RT-PCR products of patients carrying the c.520C>T showed the presence of the mutant FL transcripts but in lower amount than WT FL transcripts, indicating that c.520C>T variant causes an important but not total splicing defect (Figure S6). Furthermore, c.566A>G and c.517-1G>A variations were associated to total splicing defects. Indeed, c.566A>G is responsible of the deletion of the last 70 nt which represent around 50% of the BRCA2 transcripts in LCL in the presence of puromycin (Table 1). Sequencing of the RT-PCR products revealed the absence of the FL transcripts carrying the variation (Figure S6). The allele carrying c.517-1G>A generates two different transcripts: a minor transcript without exon 7 (Δ7 = 7%) and a predominant transcript deleted of the first nucleotide (Δ7p(1nt) = 44%) both detected by capillary electrophoresis and sequencing of the RT-PCR products (Table 1). One should note that, in the minigene-based splicing assay, the pCAS2-*BRCA2*e7 c.517-1G>A generates as well those two transcripts but in different proportion when compared to patient's RNA sample (Δ7p(1nt) = 48% and Δ7 = 52%) (Table 1 and Figure 1), suggesting that the type and severity of the splicing defect caused by this variant depends on surrounding nucleotide context and/or the cellular type. Finally, patients carrying the 7 remaining variants (c.517-20A>G, c.517-19C>T, c.538_539dup, c.587G>A, c.599C>T, c.617C>G) display an intermediate splicing defect when compared to

healthy controls and patients carrying the *BRCA2* exon 7 genomic deletion, with increase in Δ7 transcript levels to different extents, i.e. from 2 to 29% depending on the variant, showing that these variants induce minor to moderate splicing defects (Table 1). Remarkably, the severity of the partial splicing defects observed in patients RNA sample reflects that observed in the splicing minigene assays (Figure S8).

Allele specific expression analyses. In order to quantitatively evaluate the contribution of wild-type and mutant alleles to the production of *BRCA2* transcripts containing exon 7 and because Sanger sequencing is known to have low detection sensitivity, we took advantage of the quantitative nature of SNaPshot assay allowing to measure allele specific expression (ASE), i.e. the relative amount of each allele within the RT-PCR product containing exon 7. As expected, analysis of the allelic expression for patients carrying the c.517G>T, c.566A>G and c.631G>A variations indicates that the FL transcripts expressed from the mutant allele were drastically reduced or absent as compared to the WT allele (2, 0 and 3% for c.517G>T, c.566A>G and c.631G>A, respectively) (Table 1). Given the results of our minigene assays and RT-PCR analysis of patient's RNA, we conclude that this allelic imbalance is essentially due to variant-induced exon 7 skipping. Similarly, we also observed an allelic imbalance for 587G>A, c.599C>T, c.520C>T, c.538_539dup and c.617C>G, from 92% to 25%, as compared to the expression of the WT allele (Table 1), thus confirming that that theses variant induce weak to severe exon skipping. In contrast, we did not observe an allelic imbalance for the c.521G>A, c.573T>C, c.575T>C and c.627C>T (102, 99, 105 and 98% expression relative to WT, respectively), which confirms that these variations do not affect exon 7 splicing (Table 1). As of note, allele specific expression obtained here by Snapshot for c.520C>T and c.617C>G are concordant with those previously obtained by pyrosequencing on the same samples (Gaildrat *et al.*, 2012). Surprisingly, we observed an allelic imbalance for c.605C>G and c.625C>T in favour of the mutant allele (139 and 177% expression relative to WT, respectively) (Table 1). These variants are not associated to splicing defects in the minigene assays nor in patient's RNA, suggesting that another variation located on the alleged WT alleles might be responsible of their lower expression. Indeed, the patients carrying the c.605C>G and c.625C>T variations also harbour a deleterious variation located in trans in exon 11, c.5212_5216del and c.1310_1313del, respectively. Those two out-of-frame small deletions lead to the introduction of a premature termination codon (PTC) in *BRCA2* transcripts most likely targeted to degradation by the NMD pathway.

Comparison of patient RNA data with results from the ex vivo splicing reporter minigene-based assay. Altogether, the results obtained with patient samples agree with the minigene assays, and highlight the physiological pertinence of the minigene-based splicing assay for analysis of BRCA2 exon 7 variants. Nonetheless, it is important to note that the WT pCAS2-*BRCA2*e7 minigene did not fully reproduce the splicing pattern of the endogenous *BRCA2* exon 7 expressed in the control RNA samples, as the alternative skipping of *BRCA2* exon 7 observed in minigene assay (Δ7 = 13%) is very residual in patient RNA (Δ7 = 1%) (Table 1). In consequence, variants increasing exon 7 inclusion cannot be easily detected by patient RNA analysis. However, we observed a striking logarithmic correlation (R^2 = 0.98) between the severity of the splicing defects evaluated in the monoallelic minigene-based assay and those assessed in patient LCLs carrying the same variant at the heterozygous state (Figure S7). Interestingly, this correlation is logarithmic, suggesting that the minigene system might overestimate moderate partial splicing defects, such as these induced by c.517-19C>T, c.538_539dup, c.587G>A and c.599C>T.

Familial data (Sandrine Caputo, in progress).

Complementary multifactorial likelihood analyses, based on the collect of genetic, clinical, tumoral, cosegregation and familial data of patients carrying the natural variations of interest, are currently underway in order to further evaluate the pathogenic nature of this type of variation.

Discussion

Recently, we reported the effect on splicing of 32 variants mapping to *BRCA2* exon 7 outside its 3' and 5'ss, of which 11 were found to induce exon skipping probably by modifying cis-regulatory elements (Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012). Moreover, we found that these splicing defects could be predicted by a new *in silico* method based on ΔtESRseq scores (QUEPASA). Lately, we took advantage of this experimental dataset to evaluate the performance of three additional approaches aiming to predict ESR alterations (Soukarieh *et al.*, 2016; Tubeuf *et al.*, in preparation). Here, we decided to further assess the predictive power of splicing-dedicated *in silico* methods by extending our analysis to a new dataset of naturally occurring variants detected in this exon (n= 66) and in flanking intronic sequences (n=15) reported by either molecular diagnostic

laboratories within the ENIGMA consortium or by different human variation databases. Before this study, only 16 SNVs in *BRCA2* exon 7 had been reported as causing aberrant splicing, all shown to increase 7 exon skipping (Biswas *et al.*, 2011; Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012; Pensabene *et al.*, 2009; Sanz *et al.*, 2010). By resorting both to minigene-based splicing assay and patient RNA analysis, when available, our work uncovered 64 new splicing mutations bringing the number of *BRCA2* exon 7 splicing mutations to a total of 89. More importantly, we found that the majority of *BRCA2* exon 7 splicing mutations (~74%) map outside the splice sites, indicating an important contribution of variant-induced ESR alterations in this exon. Hence, our work confirms that *BRCA2* exon 7 is very sensitive to ESR-mutations and that this type of alterations are still underestimated. Our results further suggest that *BRCA2* exon 7 contains a high density of regulatory elements distributed along the exon, that contribute to its efficient splicing.

A major aim of this study was to evaluate the performance of various bioinformatics tools aiming at predicting the impact of sequence variants on RNA splicing by using the experimental output of the pCAS2-*BRCA2*e7 minigene-based splicing assay as a benchmark dataset. Given that the consensus sequences of human 3' and 5' splice sites have been thoroughly established (Cartegni *et al.*, 2002; Shapiro and Senapathy, 1987), programs devoted to predicting alterations in consensus splice sites have been extensively characterized, including on BRCA genes, and are currently the most frequently used filtering tools in variant stratification strategies for molecular diagnostic purpuses (Houdayer *et al.*, 2012; Leman *et al.*, 2018; Spurdle *et al.*, 2008; Théry *et al.*, 2011; Vreeswijk *et al.*, 2009). Consistently, in the case of *BRCA2* exon 7, our results showed a good concordance between minigene data and *in silico* results from MES, SSFL, MES+SSFL and SPiCE for variants mapping within the splice-site consensus sequence. Furthermore, among the two algorithms tested (MES and SSFL) and their combination (MES+SSFL and SPiCE), we found that SPiCE provides the best performances for predicting exon skipping-mutations, as previously suggested (Leman *et al.*, 2018). However, our results pinpoint two limits of SPiCE which cannot, at this point in time, predict the type of the splicing defects, i.e. exon skipping or use of another splice site (alternative, cryptic or *de novo*) as well as alterations in other splicing signals mapping upstream the 3'ss, such as the BP, which require specific computational tools not yet fully evaluated. As for consequences of nucleotide changes occurring at splicing regulatory elements, we evaluated the discriminating power of four computational tools (QUEPASA, HEXplorer, SPANR and HAL) recently described as promising for screening splicing regulatory mutations

(Soukarieh *et al.*, 2016). Our results confirm and extend previous work that highlighted the good reliability of these approaches used alone or in combination that accurately pinpoint variants that induce exon skipping though ESR alteration as well as the importance of the splicing defects (Soukarieh *et al.*, 2016; Tubeuf *et al.*, in preparation), further pinpointing their potential to be used as filtering tools in variant stratification strategies. Overall, our data demonstrate the need for developing a fully comprehensive program that will integrate and combine multiple splicing signals-dedicated *in silico* tools or pipeline in order to propose a decision-making tool to guide geneticists towards the identification of spliceogenic, potentially pathogenic, variants in the deluge of high-throughput sequencing data.

Importantly, among the 89 splicing mutations in *BRCA2* exon 7, we identified 35 SNVs causing increased exon inclusion. To our knowledge, this is the first report of SNVs having a positive impact on *BRCA2* splicing. A few cases of variant-induced exon inclusion have already been reported in other genes and can have a significant clinical impact, either by attenuating the severity of the disease (Prior *et al.*, 2009; Vezain *et al.*, 2010) or by leading to disease (Liu and Gong, 2008). However, in the case of *BRCA2* exon 7, it is difficult to predict the biological and clinical consequences of these splicing alterations given that we were not able to confirm variant-induced exon inclusion in patient RNA. Although a high concordance for splicing analysis was demonstrated for multiple disease-associated *BRCA1/2* variants between splicing reporter minigene assays and patient-derived lymphocyte mRNA, differences in splicing patterns have been observed for a proportion of spliceogenic variants (Acedo *et al.*, 2012; Bonnet *et al.*, 2008; Steffensen *et al.*, 2014). As we used blood derived samples as a main source of patient RNA and cervix-derived HeLa cells to express the splicing minigene reporter, we cannot exclude the existence of a tissue-specific alternative splicing program that could explain differences between minigene and patient RNA. Alternatively, it is possible that the observed differences in exon 7 splicing pattern might be due to the artificial nature of minigene constructs which lack features of the natural genomic environment, i.e. such as the full gene sequence, native promoters or chromatin structure (Baralle *et al.*, 2006). Therefore, the optimization of minigene constructs by mimicing the natural genomic context might allow to overcome this limitation (Acedo *et al.*, 2012, 2015; Raponi *et al.*, 2012; Sangermano *et al.*, 2018; Sharma *et al.*, 2014). The most suitable method to identify splicing aberrations is based on the analysis of the splicing pattern of transcripts expressed in the affected tissue of the patients as compared to the equivalent normal tissue (Baralle *et al.*, 2009a).

However, the relevant tissue is rarely available, when genes of interest are expressed in blood cells, blood samples can be used as surrogate biological systems to perform RNA splicing analyses. However, this type of samples is often unavailable as well. Therefore, minigene-based splicing assays have become an alternative approach to evaluate the consequences of DNA variants on splicing (Gaildrat *et al.,* 2010). Yet, it is important to validate, exon by exon, the minigene splicing reporter assay by extensive comparisons with patient RNA data, ideally derived from blood and relevant tissue, i.e. breast and ovary in the case of HBOC (Buratti *et al.*, 2013).

One of the major advantages of the splicing reporter minigene assay is that transcripts produced from particular minigenes such as the pCAS2 vector are not degraded by the NMD, contrary to patient RNA analysis where the NMD selectively degrades mRNAs harboring PTC that can impair the relative proportions of each isoform, leading sometimes to misinterpretation (Caminsky *et al.*, 2014). Surprisingly, we detected an important proportion of mutant PTC-containing *BRCA2* transcripts in blood-derived RNA sources, however, in higher levels in whole blood samples when compared to LCLs of the same patients. Our data thus suggest that NMD is deficient in PAXgene stabilized blood samples and incomplete in LCLs. These findings are in concordance with those previously reported for PTC-harboring *OPA1* transcripts, demonstrating a lower NMD efficiency in whole blood, associated with high levels of PTC-harboring transcripts (40-50%), as compared to LCLs (20–30%) of the same patients (Schimpf *et al.*, 2008).

Most of the time, spliceogenic variations are associated with a decreased expression of functional BRCA2 through introduction of a PTC which is likely to be targeted by the NMD and/or to lead to the production of a truncated protein and are thus considered as deleterious mutations (Walker *et al.*, 2013). While this is generally the case, occasionally multiple splice variants are generated including some that restore open-reading frame and result in a partially functional protein (Li *et al.*, 2009). In the case of BRCA2 exon 7, the supposedly pathogenic variations c.631+2T>G and c.587G>A (p.W194X) have been shown to induce mostly exon 7 skipping, but also lead to the production of a minor naturally occurring alternative transcript lacking exons 4 to 7 (*BRCA2* Δ4-7) (Biswas *et al.*, 2011). This transcript, restoring open-reading frame and bypassing the PTC, encodes a protein deleted in its internal part of 105 amino acids (BRCA2Δ105) which is proficient in homologous recombination-mediated DNA repair and in tumor suppression (Biswas *et al.*, 2011; Thirthagiri *et al.*, 2016). In addition, exons 4-7 were shown to be dispensable for viability, fertility,

normal development of mice and tumor-free survival (Thirthagiri *et al.*, 2016). Moreover, c.631+2T>G has been reported at a homozygous state in at least two children with Fanconi Anemia (Wagner *et al.*, 2004), supporting the hypomorphic nature of c.631+2T>G and suggesting that others variants affecting BRCA2 exon 7 splicing may also lead to similar phenotypes. Importantly, a recent work conducted within the ENIGMA consortium based on a case-control association study and functional analyses, has suggested that hypomorphic BRCA2 missense variant can confer an intermediate or moderate increased risk to develop breast cancer (Shimelis *et al.*, 2017). Further studies are thus required to fully establish the clinical relevance and estimate the risks of variations occurring in *BRCA2* exon 4-7.

In conclusion, our work confirm the power of combining splicing-dedicated *in silico* and functional analyses for identifying spliceogenic disease-causing candidates among the deluge of variants detected by molecular diagnostic laboratories. These data can complement further analyses at the protein level and have implications in variant interpretation.

Bibliography

Acedo, A., Sanz, D.J., Durán, M., Infante, M., Pérez-Cabornero, L., Miner, C., and Velasco, E.A. (2012). Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. Breast Cancer Res. BCR *14*, R87.

Acedo, A., Hernández-Moro, C., Curiel-García, Á., Díez-Gómez, B., and Velasco, E.A. (2015). Functional classification of BRCA2 DNA variants by splicing assays in a large minigene with 9 exons. Hum. Mutat. *36*, 210–221.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Baralle, D., Lucassen, A., and Buratti, E. (2009). Missed threads. The impact of pre-mRNA splicing defects on clinical practice. EMBO Rep. *10*, 810–816.

Baralle, M., Skoko, N., Knezevich, A., De Conti, L., Motti, D., Bhuvanagiri, M., Baralle, D., Buratti, E., and Baralle, F.E. (2006). NF1 mRNA biogenesis: effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. FEBS Lett. *580*, 4449–4456.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. BMC Bioinformatics *15*, 293.

Bateman, J.F., Freddi, S., Lamandé, S.R., Byers, P., Nasioulas, S., Douglas, J., Otway, R., Kohonen-Corish, M., Edkins, E., and Forrest, S. (1999). Reliable and sensitive detection of premature termination mutations using a protein truncation test designed to overcome problems of nonsense-mediated mRNA instability. Hum. Mutat. *13*, 311–317.

Bateman, J.F., Freddi, S., Nattrass, G., and Savarirayan, R. (2003). Tissue-specific RNA surveillance? Nonsense-mediated mRNA decay causes collagen X haploinsufficiency in Schmid metaphyseal chondrodysplasia cartilage. Hum. Mol. Genet. *12*, 217–225.

Béroud, C., Letovsky, S.I., Braastad, C.D., Caputo, S.M., Beaudoux, O., Bignon, Y.J., Bressac-De Paillerets, B., Bronner, M., Buell, C.M., Collod-Béroud, G., *et al.* (2016). BRCA Share: A Collection of Clinical BRCA Gene Variants. Hum. Mutat. *37*, 1318–1328.

Biswas, K., Das, R., Alter, B.P., Kuznetsov, S.G., Stauffer, S., North, S.L., Burkett, S., Brody, L.C., Meyer, S., Byrd, R.A., *et al.* (2011). A comprehensive functional characterization of BRCA2 variants associated with Fanconi anemia using mouse ES cell-based assay. Blood *118*, 2430–2442.

Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., Berthet, P., Chevrier, A., Dugast, C., Layet, V., *et al.* (2008). 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. *45*, 438–446.

Buratti, E., Baralle, M., and Baralle, F.E. (2013). From single splicing events to thousands: the ambiguous step forward in splicing research. Brief. Funct. Genomics *12*, 3–12.

Caminsky, N., Mucaki, E.J., and Rogan, P.K. (2014). Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Research *3*, 282.

Caputo, S., Benboudjema, L., Sinilnikova, O., Rouleau, E., Béroud, C., Lidereau, R., and French BRCA GGC Consortium (2012). Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. Nucleic Acids Res. *40*, D992-1002.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Corvelo, A., Hallegger, M., Smith, C.W.J., and Eyras, E. (2010). Genome-wide association between branch point properties and alternative splicing. PLoS Comput. Biol. *6*, e1001016.

Davy, G., Rousselin, A., Goardon, N., Castéra, L., Harter, V., Legros, A., Muller, E., Fouillet, R., Brault, B., Smirnova, A.S., *et al.* (2017). Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. Eur. J. Hum. Genet. EJHG *25*, 1147–1154.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C.

(2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., and Martins, A. (2013). 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. *34*, 1547–1557.

Eccles, D.M., Mitchell, G., Monteiro, A.N.A., Schmutzler, R., Couch, F.J., Spurdle, A.B., Gómez-García, E.B., and ENIGMA Clinical Working Group (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. *26*, 2057–2065.

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J.O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res. *42*, 10681–10697.

Fackenthal, J.D., Yoshimatsu, T., Zhang, B., de Garibay, G.R., Colombo, M., De Vecchi, G., Ayoub, S.C., Lal, K., Olopade, O.I., Vega, A., *et al.* (2016). Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples. J. Med. Genet. *53*, 548–558.

Fokkema, I.F.A.C., Taschner, P.E.M., Schaafsma, G.C.P., Celli, J., Laros, J.F.J., and den Dunnen, J.T. (2011). LOVD v.2.0: the next generation in gene variant databases. Hum. Mutat. *32*, 557–563.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., *et al.* (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Frebourg, T. (2014). The challenge for the next generation of medical geneticists. Hum. Mutat. *35*, 909–911.

Freddi, S., Savarirayan, R., and Bateman, J.F. (2000). Molecular diagnosis of Stickler syndrome: a COL2A1 stop codon mutation screening strategy that is not compromised by mutant mRNA instability. Am. J. Med. Genet. *90*, 398–406.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. Clifton NJ *653*, 249–257.

Gaildrat, P., Krieger, S., Di Giacomo, D., Abdat, J., Révillion, F., Caputo, S., Vaur, D., Jamard, E., Bohers, E., Ledemeney, D., *et al.* (2012). Multiple sequence variants of BRCA2 exon 7 alter splicing regulation. J. Med. Genet. *49*, 609–617.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res. *36*, 2257–2267.

Guidugli, L., Carreira, A., Caputo, S.M., Ehlen, A., Galli, A., Monteiro, A.N.A., Neuhausen, S.L., Hansen, T.V.O., Couch, F.J., Vreeswijk, M.P.G., *et al.* (2014). Functional assays for analysis of

variants of uncertain significance in BRCA2. Hum. Mutat. *35*, 151–164.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., and Pease, L.R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene *77*, 51–59.

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., *et al.* (2012). 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. *33*, 1228–1238.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kol, G., Lev-Maor, G., and Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. Hum. Mol. Genet. *14*, 1559–1568.

Lamandé, S.R., Bateman, J.F., Hutchison, W., McKinlay Gardner, R.J., Bower, S.P., Byrne, E., and Dahl, H.H. (1998). Reduced collagen VI causes Bethlem myopathy: a heterozygous COL6A1 nonsense mutation results in mRNA decay and functional haploinsufficiency. Hum. Mol. Genet. *7*, 981–989.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980-985.

Ledermann, J., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., Scott, C.L., Meier, W., Shapira-Frommer, R., Safra, T., *et al.* (2014). Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. Lancet Oncol. *15*, 852–861.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Leman, R., Gaildrat, P., Gac, G.L., Ka, C., Fichou, Y., Audrezet, M.-P., Caux-Moncoutier, V., Caputo, S.M., Boutry-Kryza, N., Léone, M., *et al.* (2018). 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.

Li, L., Biswas, K., Habib, L.A., Kuznetsov, S.G., Hamel, N., Kirchhoff, T., Wong, N., Armel, S., Chong, G., Narod, S.A., *et al.* (2009). Functional redundancy of exon 12 of BRCA2 revealed by a comprehensive analysis of the c.6853A>G (p.I2285V) variant. Hum. Mutat. *30*, 1543–1550.

Lindor, N.M., Guidugli, L., Wang, X., Vallée, M.P., Monteiro, A.N.A., Tavtigian, S., Goldgar, D.E., and Couch, F.J. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). Hum. Mutat. *33*, 8–21.

Liu, F., and Gong, C.-X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol. Neurodegener. *3*, 8.

Magyar, I., Colman, D., Arnold, E., Baumgartner, D., Bottani, A., Fokstuen, S., Addor, M.-C., Berger, W., Carrel, T., Steinmann, B., *et al.* (2009). Quantitative sequence analysis of FBN1 premature termination codons provides evidence for incomplete NMD in leukocytes. Hum. Mutat. *30*, 1355–1364.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res. *25*, 290–303.

Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., Santamariña, M., Behar, R., Blanco, A., Carrasco, E., *et al.* (2018). Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? Front. Genet. *9*, 366.

Motohashi, K. (2015). A simple and efficient seamless DNA cloning method using SLiCE from Escherichia coli laboratory strains and its application to SLiP site-directed mutagenesis. BMC Biotechnol. *15*, 47.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Pensabene, M., Spagnoletti, I., Capuano, I., Condello, C., Pepe, S., Contegiacomo, A., Lombardi, G., Bevilacqua, G., and Caligo, M.A. (2009). Two mutations of BRCA2 gene at exon and splicing site in a woman who underwent oncogenetic counseling. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. *20*, 874–878.

Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. *32*, 577–591.

Plon, S.E., Eccles, D.M., Easton, D., Foulkes, W.D., Genuardi, M., Greenblatt, M.S., Hogervorst, F.B.L., Hoogerbrugge, N., Spurdle, A.B., Tavtigian, S.V., *et al.* (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum. Mutat. *29*, 1282–1291.

Prior, T.W., Krainer, A.R., Hua, Y., Swoboda, K.J., Snyder, P.C., Bridgeman, S.J., Burghes, A.H.M., and Kissel, J.T. (2009). A positive modifier of spinal muscular atrophy in the SMN2 gene. Am. J. Hum. Genet. *85*, 408–413.

Raponi, M., Douglas, A.G.L., Tammaro, C., Wilson, D.I., and Baralle, D. (2012). Evolutionary constraint helps unmask a splicing regulatory region in BRCA1 exon 11. PloS One *7*, e37255.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., *et al.* (2015). 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. *17*, 405–424.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Sangermano, R., Khan, M., Cornelis, S.S., Richelle, V., Albert, S., Garanto, A., Elmelik, D., Qamar, R., Lugtenberg, D., van den Born, L.I., *et al.* (2018). ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease. Genome Res. *28*, 100–110.

Sanz, D.J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardeñosa, E., Lastra, E., Pagani, F., Miner, C., and Velasco, E.A. (2010). A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *16*, 1957–1967.

Schimpf, S., Fuhrmann, N., Schaich, S., and Wissinger, B. (2008). Comprehensive cDNA study and quantitative transcript analysis of mutant OPA1 transcripts containing premature termination codons. Hum. Mutat. *29*, 106–112.

Schwartz, S., Hall, E., and Ast, G. (2009). SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. Nucleic Acids Res. *37*, W189-192.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155–7174.

Sharma, N., Sosnay, P.R., Ramalho, A.S., Douville, C., Franca, A., Gottschalk, L.B., Park, J., Lee, M., Vecchio-Pagan, B., Raraigh, K.S., *et al.* (2014). Experimental assessment of splicing variants using expression minigenes and comparison with *in silico* predictions. Hum. Mutat. *35*, 1249–1259.

Shimelis, H., Mesman, R.L.S., Von Nicolai, C., Ehlen, A., Guidugli, L., Martin, C., Calléja, F.M.G.R., Meeks, H., Hallberg, E., Hinton, J., *et al.* (2017). BRCA2 Hypomorphic Missense Variants Confer Moderate Risks of Breast Cancer. Cancer Res. *77*, 2789–2799.

Signal, B., Gloss, B.S., Dinger, M.E., and Mercer, T.R. (2018). Machine learning annotation of human branchpoints. Bioinforma. Oxf. Engl. *34*, 920–927.

Sonnenblick, A., de Azambuja, E., Azim, H.A., and Piccart, M. (2015). An update on PARP inhibitors--moving to the adjuvant setting. Nat. Rev. Clin. Oncol. *12*, 27–41.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., Tosi, M., and Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using *In silico* Tools. PLoS Genet. *12*, e1005756.

Spurdle, A.B., Couch, F.J., Hogervorst, F.B.L., Radice, P., Sinilnikova, O.M., and IARC Unclassified Genetic Variants Working Group (2008). Prediction and assessment of splicing alterations: implications for clinical testing. Hum. Mutat. *29*, 1304–1313.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T. vO (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. Eur. J. Hum. Genet. EJHG *22*, 1362–1368.

Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum. Genet. *136*, 665–677.

Szabo, C., Masiello, A., Ryan, J.F., and Brody, L.C. (2000). The breast cancer information core: database design, structure, and scope. Hum. Mutat. *16*, 123–131.

Tavtigian, S.V., Deffenbaugh, A.M., Yin, L., Judkins, T., Scholl, T., Samollow, P.B., de Silva, D., Zharkikh, A., and Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J. Med. Genet. *43*, 295–305.

Tavtigian, S.V., Byrnes, G.B., Goldgar, D.E., and Thomas, A. (2008). Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. Hum. Mutat. *29*, 1342–1354.

Théry, J.C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M.-P., Killian, A., Duponchel, C., Rousselin, A., Vaur, D., Peyrat, J.-P., *et al.* (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur. J. Hum. Genet. *19*, 1052–1058.

Thirthagiri, E., Klarmann, K.D., Shukla, A.K., Southon, E., Biswas, K., Martin, B.K., North, S.L., Magidson, V., Burkett, S., Haines, D.C., *et al.* (2016). BRCA2 minor transcript lacking exons 4-7 supports viability in mice and may account for survival of humans with a pathogenic biallelic mutation. Hum. Mol. Genet. *25*, 1934–1945.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J., *et al.* (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. *29*, 1412–1424.

Vallée, M.P., Di Sera, T.L., Nix, D.A., Paquette, A.M., Parsons, M.T., Bell, R., Hoffman, A., Hogervorst, F.B.L., Goldgar, D.E., Spurdle, A.B., *et al.* (2016). Adding *In silico* Assessment of Potential Splice Aberration to the Integrated Evaluation of BRCA Gene Unclassified Variants. Hum. Mutat. *37*, 627–639.

Vezain, M., Saugier-Veber, P., Goina, E., Touraine, R., Manel, V., Toutain, A., Fehrenbach, S., Frébourg, T., Pagani, F., Tosi, M., *et al.* (2010). A rare SMN2 variant in a previously unrecognized

composite splicing regulatory element induces exon 7 inclusion and reduces the clinical severity of spinal muscular atrophy. Hum. Mutat. *31*, E1110-1125.

Vreeswijk, M.P.G., Kraan, J.N., van der Klift, H.M., Vink, G.R., Cornelisse, C.J., Wijnen, J.T., Bakker, E., van Asperen, C.J., and Devilee, P. (2009). Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. Hum. Mutat. *30*, 107–114.

Wagner, J.E., Tolar, J., Levran, O., Scholl, T., Deffenbaugh, A., Satagopan, J., Ben-Porat, L., Mah, K., Batish, S.D., Kutler, D.I., *et al.* (2004). Germline mutations in BRCA2: shared genetic susceptibility to breast cancer, early onset leukemia, and Fanconi anemia. Blood *103*, 3226–3229.

Walker, L.C., Whiley, P.J., Houdayer, C., Hansen, T.V.O., Vega, A., Santamarina, M., Blanco, A., Fachal, L., Southey, M.C., Lafferty, A., *et al.* (2013). 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. *34*, 1424–1431.

Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. *8*, 749–761.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. J. Comput. Mol. Cell Biol. *11*, 377–394.

Zhang, Q., Fan, X., Wang, Y., Sun, M.-A., Shao, J., and Guo, D. (2017). BPP: a sequence-based algorithm for branch point prediction. Bioinforma. Oxf. Engl. *33*, 3166–3172.

Legends to figures and tables.

Figure 1. A relatively important fraction of intronic variants alters the splicing pattern of *BRCA2* exon 7 in agreement with *in silico* predictions. (A) Distribution of the 15 natural intronic variants reported in the vincity of *BRCA2* exon 7 splice sites. The diagram shows the nucleotide composition of *BRCA* exon 7 and its flanking intronic regions (c.517-30_c.517+30), the amino-acid sequence encoded by exon 7 (1-letter code, p.173_p.211), as well as the relative position and identity of each variant. (B) RT-PCR analysis of the splicing pattern of pCAS2-*BRCA2*e7 minigenes carrying the variants of interest. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ±

SEM of three independent transfection experiments. The identities of the RT-PCR products are indicated on the right. (C) Comparison of the performances of splice site-dedicated bioinformatics approaches in predicting variant-induced alterations in *BRCA2* exon 7 3'ss and 5'ss. Results are based on the comparison of the experimental data obtained in B with the *in silico* evaluation presented in Table S5. *In silico* predictions of potential splice site alterations were conducted by using MES and SSFL as well as MES+SSFL and SPiCE, as described under Materials and Methods. True and false calls of variant-induced exon-skipping were determined by taking into account the following thresholds: -15% for MES, -5% for SSFL (as suggested in Houdayer *et al.*, 2012) and 11,5% for SPiCE (as suggested in Leman *et al.*, 2018).

Figure 2. *BRCA2* intron 6 contains multiple branch points, as predicted by bioinformatics approaches, which are required for efficient splicing of exon 7. (A) RT-PCR analysis of the splicing pattern of pCAS2-*BRCA2*e7 minigenes carrying variants potentially disrupting putative branch points (BPs). The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products obtained are indicated on the right. (b) Identification by *in silico* predictions tools of potential *BRCA2* exon 7 BPs. The diagram shows the nucleotide composition of *BRCA2* intron 6 (c.517-65-c.517-1) as well as the relative position and scores of the putative BPs predicted by using Branchpointer, BPP, HSF, SVM-BP finder and SROOGLE (a, scores based on Kol *et al.*, 2005 and b, scores based on Schwartz *et al.*, 2008). (C) Strategy to experimentally map the branch point of *BRCA2* exon 7. The image represents the lariat protected from RNase R digestion at the 5'ss-BP junction. Arrows indicate the primers used during the first and the second step of the nested RT-PCR reactions performed on endogenous transcripts of a human lymphoblastoid cell line (F1 and R1, and F2 and R2, respectively) as described under Materials and Methods. (D) DNA sequencing electropherograms of the nested RT-PCR product corresponding to the *BRCA2* intron 6 lariats showing the exact position of the branch points (BP, black triangle). Sequencing was performed with the F2 primer. Detecting a A-to-T mismatch at the BP upon sequencing the lariat can be considered as a sign of a correctly inferred branch point given that reverse transcriptases often incorporate an incorrect nucleotide at the 5' splice site-BP junction (Pineda and Bradley, 2018).

Figure 3. A staggering number of variants mapping *BRCA2* exon 7 induce splicing defects in agreement with SRE-dedicated *in silico* predictions. (A) Distribution of the 92 natural variants reported within *BRCA2* exon 7. The diagram shows the nucleotide composition of *BRCA2* exon 7, the amino-acid sequence encoded by exon 7 (1-letter code, p.173_p.211), as well as the relative position and identity of each variant. The top and the bottom panels represent variations analyzed in previous studies or this study, respectively. (C) Summary of the performances of SRE-dedicated bioinformatics approaches (QUEPASA, HEXplorer, SPANR and HAL as well as QUEPASA&HAL, AT LEAST 3 and LR_{skip}) in predicting variant-induced *BRCA2* exon 7 skipping. Results are based on the comparison of the experimental data described in Table S4 with the *in silico* predictions presented in Table S6. True and false calls were determined by taking into account the following thresholds as previously described (Tubeuf et al, in preparation): -0.5 for QUEPASA, -14 for HEXplorer, -0.1% for SPANR, -3.4% for HAL and 31.1% for LR_{skip}. (D) Comparison, by using a Venn diagram, of the false calls produced by QUEPASA, HEXplorer, SPANR and HAL on one hand, and QUEPASA&HAL, AT LEAST 3 and LR_{skip} on the other hand, in predicting variant-induced skipping of *BRCA2* exon 7.

Figure 4. Mapping of short exonic regions by serial deletions revealed a relatively important density of splicing regulatory elements within *BRCA2* exon 7. (A) Exon-walking strategy by performing serial deletions. The diagram shows the nucleotide composition of BRCA2 exon 7 (c.517-c.631). The panel above *BRCA2* exon 7 sequence represent 10-nt serial deletions, analyzed in the pCAS2-BRCA2e7 minigene assay. The panel below *BRCA2* exon 7 represent the position and identity of each BRCA2 exon 7 exonic variations potentially affecting ESR. The color of the fragments (B) and the mutations (Table S3) reflect the results obtained in the pCAS2-BRCA2e7 minigene assay relative to entire exon splicing: black, increased exon skipping; dark grey, no effect; light grey, increased exon inclusion. (B) RT-PCR analysis of the splicing pattern of pCAS2-*BRCA2*e7 minigenes carrying the 10-nt serial deletion by RT-PCR. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products are indicated on the right.

Table 1. Description of splicing outcomes of *BRCA2* exons 7 variants observed in patient's RNA samples.

[1] For each splicing functional assay, the relative quantification of splicing events was evaluated by fluorescent RT-PCR followed by capillary electrophoresis. Results represent the mean of exon 7 inclusion level ± SEM of three independent experiments. For RT-PCR analyses of patients' RNA derived from blood-derived cultured cells, the relative quantification of splicing events was performed with/without Puromycin (top/bottom value), a nonsense mediated mRNA decay pathway (NMD) inhibitor.

[2] The *BRCA2* exon 7 splicing pattern was analysed in RNA samples isolated from lymphoblastoid cell lines (LCLs, α), whole blood (LEU, β), short-term cultured peripheral blood lymphocytes (STCLs, γ) and primary cultures of stimulated peripheral blood lymphocytes (PBLs, δ) from HBOC patients and healthy control individuals treated or not with puromycin, a NMD inhibitor (Table S2). Full description of RNA samples is described in Table S4.

[3] Previous RT-PCR analyses of patients' RNA are specified when available (Gaildrat *et al.*, 2012[1]; Houdayer *et al.*, 2012[2]; Thery *et al.*, 2011[3])

Figure S1. Structure of the pCAS2-*BRCA2*e7 minigene used in the splicing reporter assay. The pCAS2-*BRCA2*e7 minigenes were generated by inserting a genomic fragment containing *BRCA2* exons 7 as well as upstream/downstream flanking intronic sequences (216 and 207 nucleotides, respectively) into the intron of the pCAS2 vector, as indicated. The pCAS2 vector was described previously (Soukarieh *et al.*, 2016) and carries two exons (A and B) with a sequence derived from the human *SERPING1/C1NH* gene, separated by an intron containing BamHI and MluI cloning sites. Boxes represent exons, and lines in between indicate introns, whereas the bent arrow specifies the cytomegalovirus (CMV) promoter and the black circle indicates the polyadenylation signal (Poly A). Arrows below the exons represent primers used in RT-PCR reactions. The star in the forward primer symbolizes a 6-FAM 5' fluorescent modification for detection of the RT-PCR products upon capillary electrophoresis.

Figure S2. Splice-site dedicated bioinformatics predictions of variant-induced creation of de novo splice sites or activation of cryptic splice sites experimentally detected in the minigene assays. (A) RT-PCR analysis of the splicing pattern of pCAS2-*BRCA2*e7 minigenes carrying the

variants associated to the use/activation of a de novo/cryptic splice site by RT-PCR. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products as well as splicing events underlying the production of the truncated (a, b, c, d, and e) RT-PCR products are indicated in B. (B) The possibility of variant induced de novo splice sites/activation of cryptic splice sites was assessed by annotating all increments in local MaxEntScan and SpliceSiteFinder-like scores. (C) Comparison of the experimental data described in A with the *in silico* predictions presented in B. Variant induced de novo splice sites/activation of cryptic splice sites were (i) correctly predicted if scores were equal or higher to those of the corresponding reference splice site (++), (ii) partially predicted if scores were lower to those of the corresponding reference splice site (+) and (iii) not predicted if no scores are associated to the cryptic site used, or alternatively, if no change in scores is observed in the mutated context as compared to the WT **(-).**

Figure S3. Comparison of exonic variant-associated splicing effects observed in the pCAS2-*BRCA2*e7 minigenes and associated SRE-dedicated *in silico* predictions. The variations were separated into 3 groups according to their impact on splicing as experimentally determined in the pCAS2-*BRCA2*e7 minigene assays and described in Supplementary Table S6. Panels A to E compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL and LR_{skip}, respectively (Supplementary Table S5). The dashed lines indicate the thresholds used in this study as shown in Tables S6. Two-sided p-values were calculated by using ANOVA or Kruskal-Wallis, as indicated in Supplementary Table S3.

Figure S4. Correlation between variant-associated exon skipping levels described in the context of the pCAS2-*BRCA2*e7 minigene assays and *in silico* data obtained with SRE-dedicated approaches. Exon skipping levels refer to semi-quantitative data obtained from the pCAS2-*BRCA2*e7 minigene assay. Panels A to E compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL and LR_{skip}, respectively (Supplementary Table S6). Determination coefficients (R^2) and two-sided p-values were determined by performing a Pearson or Spearman correlation analysis, as indicated in Supplementary Table S3.

Figure S5. Receiver operating characteristic (ROC) curve of SRE-dedicated bioinformatics approach in predicting variant-induced exon skipping. ROC curves were performed on *BRCA2* exon 7 variants (n= 92) located outside the splice sites and separated into 2 groups according to the pCAS2-*BRCA2*e7 minigene results (variants that increased exon skipping, n= 35, and those that did not, n= 57) and confronted to *in silico* predictions generated by the each bioinformatics approach. Area under the curve (AUC) is provided at 99% CI.

Figure S6. Sanger sequencing of the PCR and RT-PCR products of patients carrying variation causing near-total to total splicing defects. Partial sequence chromatograms of the gDNA (left) and cDNA (right) amplified fragment showing the presence of the mutation of interest (indicated by an arrow) in PCR products but not (or in very low proportion) in the RT-PCR products.

Figure S7. Correlation between *BRCA2* exon 7 skipping levels observed in pCAS2-*BRCA2*e7 minigene assays and patient-derived lymphoblastoid cell lines. The impact on splicing of *BRCA2* exon 7 variants was determined in the context of the pCAS2-*BRCA2*e7 minigene and in patient-derived lymphoblastoid cell lines treated with puromycin when available. The precise correspondence between the level of exon skipping observed in the pCAS2-*BRCA2*e7 minigene assays and in patient-derived lymphoblastoid cell lines treated with puromycine and the identity of the corresponding *BRCA2* exon 7 variant, is indicated on Tables 1 and S4. Coefficient of determination (R^2) was determined by performing a Spearman correlation analysis.

Table S1. Description of the primers used in this study.

[1] F, forward; R, reverse.

[2] Intronic and exonic sequences are indicated in grey and black, respectively. The position of the nucleotide variation is underlined. The double underlined sequences correspond to restriction sites for BamHI and MluI and the sequence highlighted in grey correspond to the 15bp-tail used for homologous recombination.

Table S2. Description of protocols used to analyse patient RNA samples. All patient blood-derived samples were collected within the ENIGMA consortium by the molecular diagnostic

laboratories. Then, total RNA was extracted, with standard methods as indicated from blood-derived samples or from blood-derived cultured cells treated/untreated with Puromycin (Puro +/− experiments), a nonsense mediated mRNA decay pathway (NMD) inhibitor. In all cases, RNA samples were treated with RNase-Free DNase previous to Reverse Transcription (RT) reactions. The splicing pattern of BRCA2 transcripts was analyzed by semi-quantitative fluorescent (RT-)PCR by using a combination of specific primers located in *BRCA2* exons 2 and exon 5 (Table S1) and different commercially available kits in combination with oligodT and/or random hexamers. Then, RT-PCR products were separated by electrophoresis on a 2% agarose gel, gel-purified and sequenced and splicing events were quantitated by performing capillary electrophoresis on an automated sequencer. [1] According to manufacturers' protocol.

Table S3. Description of statistical analyses conducted in this study. Data derived from confrontation of experimental and *in silico* analyses were compared by using either Student's test, one-way ANOVA test and Pearson's correlation or their derivatives depending on the purpose of the analysis and data distribution patterns.

Table S4. Description of *BRCA2* exon 7 variants selected in this study. All SNVs identified spanning *BRCA2* c.517-25 to 631+208 position in HBOC patient within the ENIGMA consortium were collected. The selection was then extended to all exonic SNVs reported in *BRCA2* exon 7 by interrogating 9 databases (BRCA-Share, BIC, ClinVar, ESP, HGMD, LOVD, COSMIC, dbSNP and gnomAD). The star indicated variants tested in minigene assay in previous studies.

[1] Variant classification is indicated when attributed and was retrieved from each database and refers to the 5-tier system used by the InSiGHT Variant Interpretation Committee (http://insight-group.org/variants/classifications/) as follows: 1, not pathogenic or begnign; 2, likely not pathogenic or likely begnign; 3, uncertain significance (also called VUS for variants of unknown significance); 4, likely pathogenic; 5, pathogenic. n/a, not available.

[2] For each splicing functional assay, the relative quantification of splicing events was evaluated by fluorescent RT-PCR followed by capillary electrophoresis. Results represent the mean of exon 7 inclusion level ± SEM of three independent experiments. For RT-PCR analyses of patients' RNA derived from blood-derived cultured cells, the relative quantification of splicing events was

performed with/without Puromycin (top/bottom value), a nonsense mediated mRNA decay pathway (NMD) inhibitor.

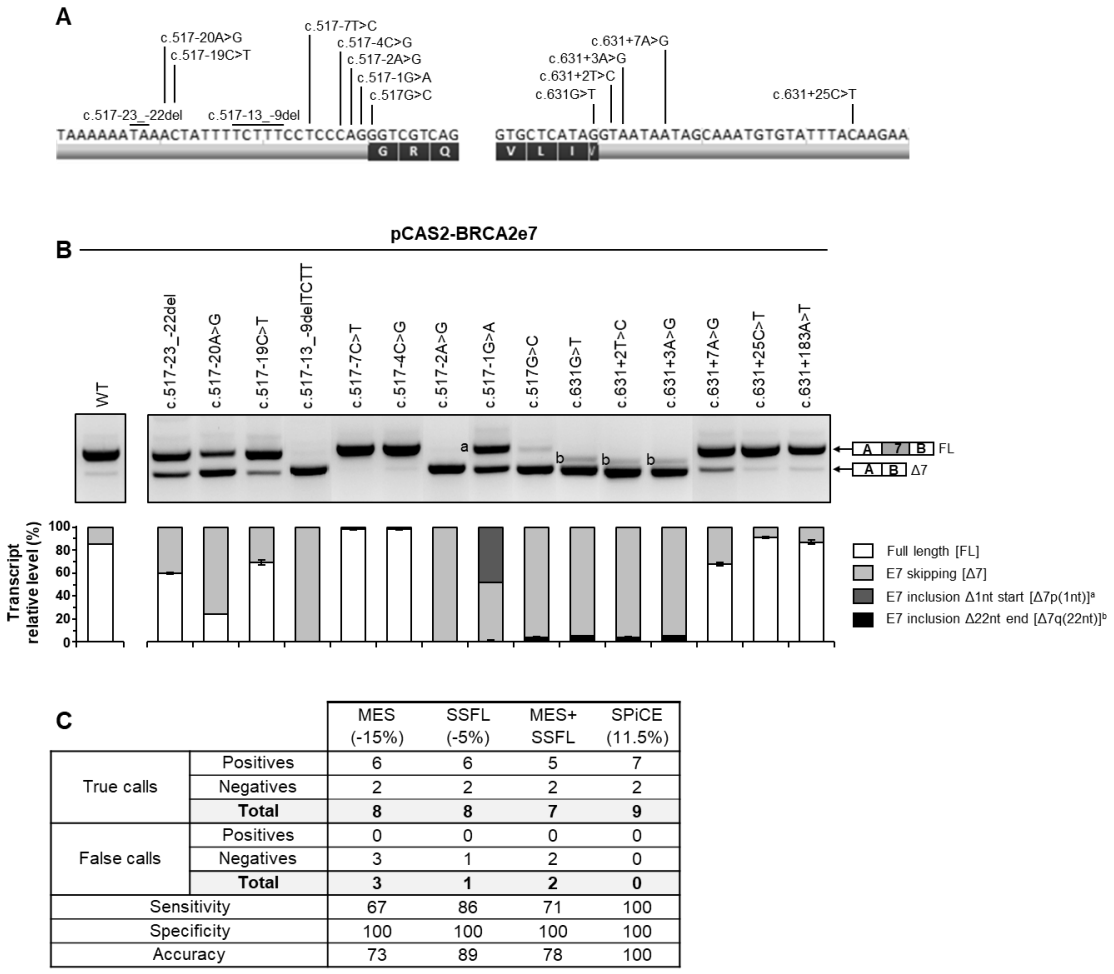
[3] Variants producing exon inclusion levels greater (increased exon inclusion) or lower (increased exon skipping) than those of the WT ± 3x SEM (15 ± 5% exon skipping) were considered as splicing mutations. The severity (partial or total) of the splicing defects induced by the variation of interest are indicated, as well as the nature of the splicing defects (+E7, increased exon inclusion; Δ7, increased exon skipping; E7Δp(1nt), deletion of the first nucleotide; E7Δq(70nt), deletion of the last 70 nt). When the variation of interest induce several splicing defects, the identities and proportion of the abnormal RT-PCR products are specified.

[4] Splicing-dedicated functional analysis (patients' RNA or minigene assay) are specified when available (Gaildrat *et al.*, 2012[1]; Di Giacomo *et al.*, 2013[2]; Rodriguez-Balada *et al.*, 2016[3]; Houdayer *et al.*, 2012[4]; Thery *et al.*, 2011[4]).

Table S5. Comparison of variant-associated splicing effects obtained with pCAS2-BRCA2e7 minigenes carrying variants mapping to BRCA2 exon 7 splice sites or to flanking intronic positions and associated splice site-dedicated *in silico* predictions. The impact on splicing of 17 variants located in or near the splice sites of *BRCA2* exon 7 was determined by performing a cell-based splicing assay with pCAS2-*BRCA2*e7 minigenes or patient RNA analysis, when available. Δ7p(1nt), inclusion of exon 7 deleted of its first nt. The table shows a separation of the variants into 2 groups according to the minigene results shown in Figure 1: variants that induced splicing defects (n=11) and those that did not (n=6). *In silico* predictions of potential effects on splicing were conducted by using 2 splice site-dedicated *in silico* tools (MES, SSFL), as well as two approaches based on their combination (MES+SSFL and SPiCE). MES and SSFL results are presented as the change in scores (Δ) of the variants relative to WT (ΔMES and ΔSSFL, respectively). True and false calls (in grey) of exon 7 splicing defects were determined by taking into account the following thresholds: -15% for ΔMES, -5% for ΔSSFL and 11.5% for SPiCE as previously recommended (Houdayer *et al.*, 2012; Leman *et al.*, 2018) and indicated between parenthesis in the Table.

Table S6. Comparison of variant-associated splicing effects obtained in the pCAS2-*BRCA2*e7 minigene assay and associated SRE-dedicated *in silico* predictions. The impact on splicing of 63 variants mapping *BRCA2* exon 7 was determined by performing a cell-based splicing assay with pCAS2-BRCA2e7 minigenes as shown in Figure S3. These variants were retrieved from human variation databases and none map reference splice sites. The table shows a separation of the variants into 3 groups according to the minigene results shown in Figure S3: variants that increased exon 7 skipping (n=23), variations with no effect on exon 7 splicing (n = 15) and those that increased exon 7 inclusion (n = 25). *In silico* predictions of potential effects on ESRs were conducted by using the 4 new SRE-dedicated *in silico* tools (QUEPASA, HEXplorer, SPANR and HAL), as well as three approaches resulting from their combination (QUEPASA&HAL, at AT LEAST 3 and LR_{skip}). True and false calls (in grey) for prediction of induced exon-skipping events were determined by taking into account the following thresholds as previously recommended for analysis of *BRCA2* exon 7 variants (Tubeuf *et al.*, in preparation) and indicated between parenthesis in the Table: -0.65 for QUEPASA (ΔtESRseq scores), -28 for HEXplorer (ΔHZEI scores), +0.35% for SPANR (Δψ scores), -5.0% for HAL (Δψ scores) and 33.6% for LR_{skip}. n/a, not applicable.

Table S7. Comparison of variant-associated splicing effects obtained in the pCAS2-*BRCA2*e7 minigene assay and associated SRE-dedicated *in silico* predictions after optimization of the thresholds. The impact on splicing of 97 *BRCA2* exon 7 variants was determined by performing a cell-based splicing assay with pCAS2-*BRCA2*e7 minigenes. The table shows a separation of the variants into 3 groups according to the minigene results shown in Table S3: variants that increased exon 7 skipping (n= 35), variations with no effect on exon 7 splicing (n = 27) and those that increased exon 7 inclusion (n = 35). *In silico* predictions of potential effects on splicing were conducted by using 4 newly developed SRE-dedicated *in silico* tools (QUEPASA, HEXplorer, SPANR and HAL), as well as three approaches resulting from their combination (QUEPASA&HAL, at least 3 and LReg). True and false calls (in grey) for prediction of exon-skipping events were determined by taking into account the following optimal thresholds determined from ROC curve analysis: -0.33 for QUEPASA (ΔtESRseq scores), -11 for HEXplorer (ΔHZEI scores), +0.38% for SPANR (Δψ scores), -4.2% for HAL (Δψ scores) and 37.5% for LR_{skip}. n/a, not applicable.



C

		MES (-15%)	SSFL (-5%)	MES+ SSFL	SPiCE (11.5%)
True calls	Positives	6	6	5	7
	Negatives	2	2	2	2
	Total	**8**	**8**	**7**	**9**
False calls	Positives	0	0	0	0
	Negatives	3	1	2	0
	Total	**3**	**1**	**2**	**0**
Sensitivity		67	86	71	100
Specificity		100	100	100	100
Accuracy		73	89	78	100

Figure 1. A relatively important fraction of intronic variants alters the splicing pattern of *BRCA2* exon 7 in agreement with *in silico* predictions.

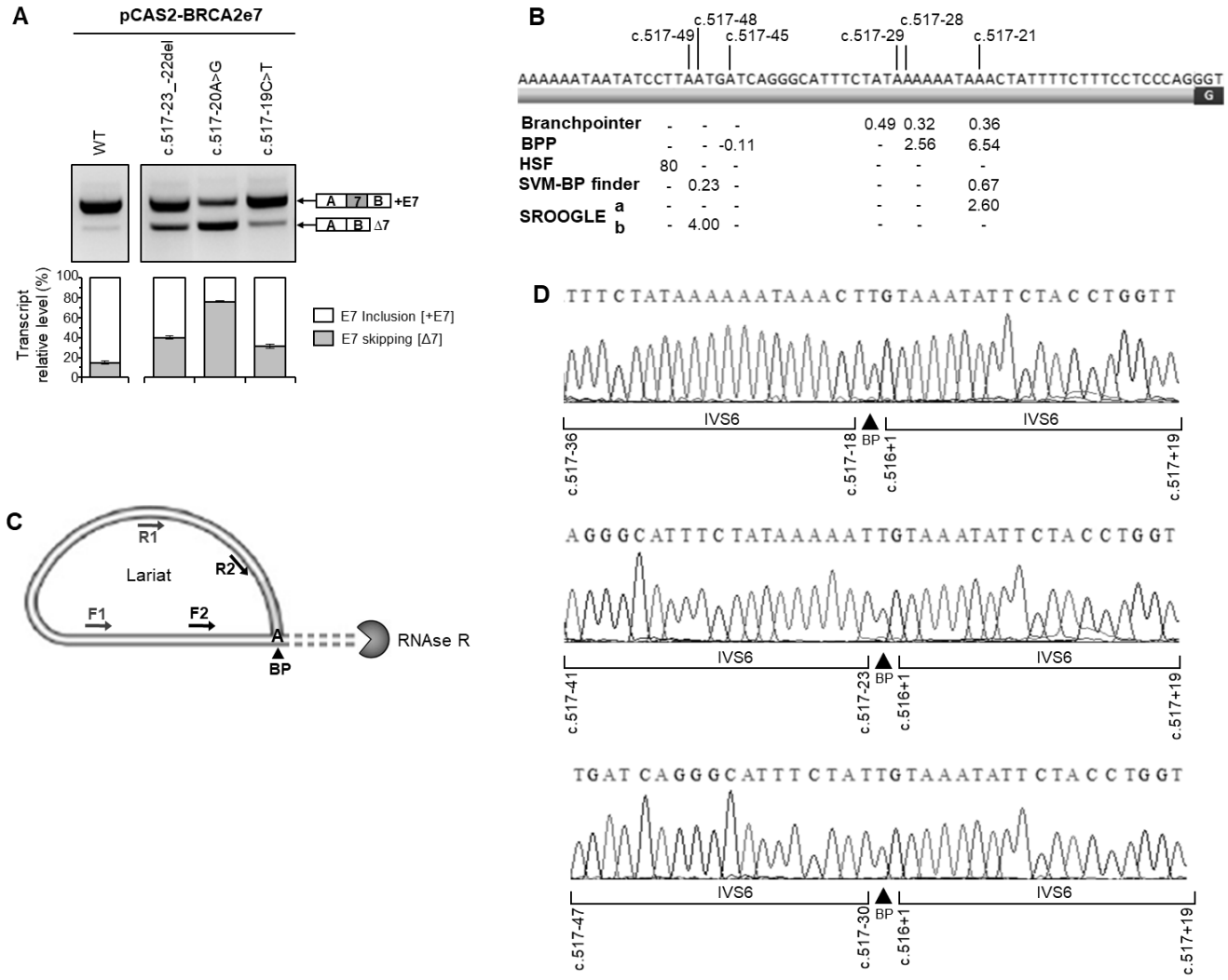
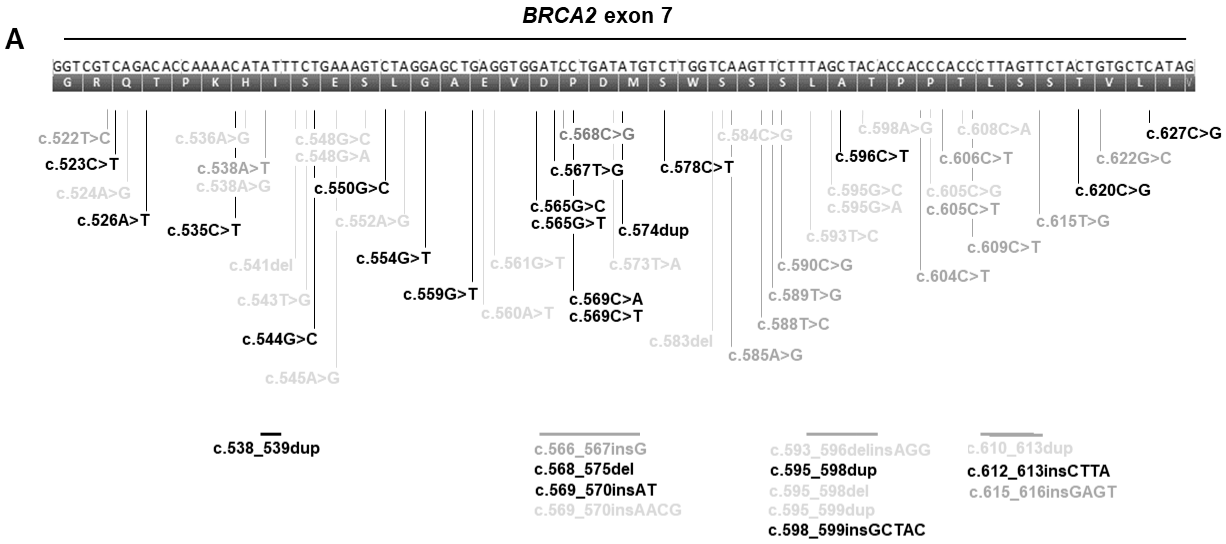


Figure 2. *BRCA2* intron 6 contains multiple branch points, as predicted by bioinformatics approaches, which are required for efficient splicing of exon 7.

A



B

		QUEPASA (-0.55)	HEXplorer (-28)	SPANR (+0,35%)	HAL (-5.0%)	LR_{skip} (33.6%)	QUEPASA & HAL	At least 3
True calls	Positives	17	11	11	19	17	17	12
	Negatives	37	37	19	32	38	39	38
	Total	**54**	**48**	**30**	**51**	**55**	**56**	**50**
False calls	Positives	3	3	12	8	2	1	2
	Negatives	6	12	5	4	6	6	11
	Total	**9**	**15**	**17**	**12**	**8**	**7**	**13**
Sensitivity (%)		74	47	69	83	74	74	52
Specificity (%)		93	93	61	80	95	98	95
Accuracy (%)		86	76	64	81	87	89	79

C

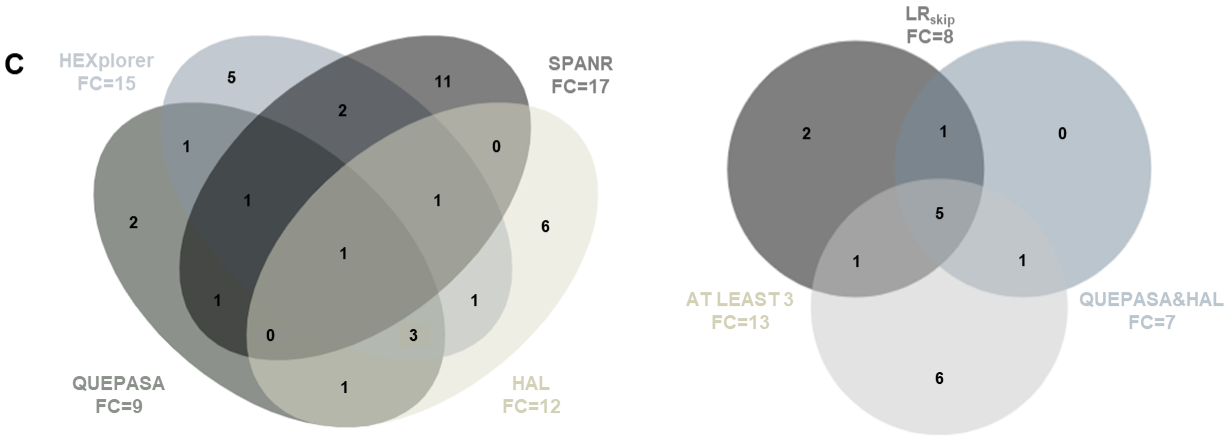


Figure 3. A staggering number of variants mapping *BRCA2* exon 7 induce splicing defects in agreement with SRE-dedicated *in silico* predictions.

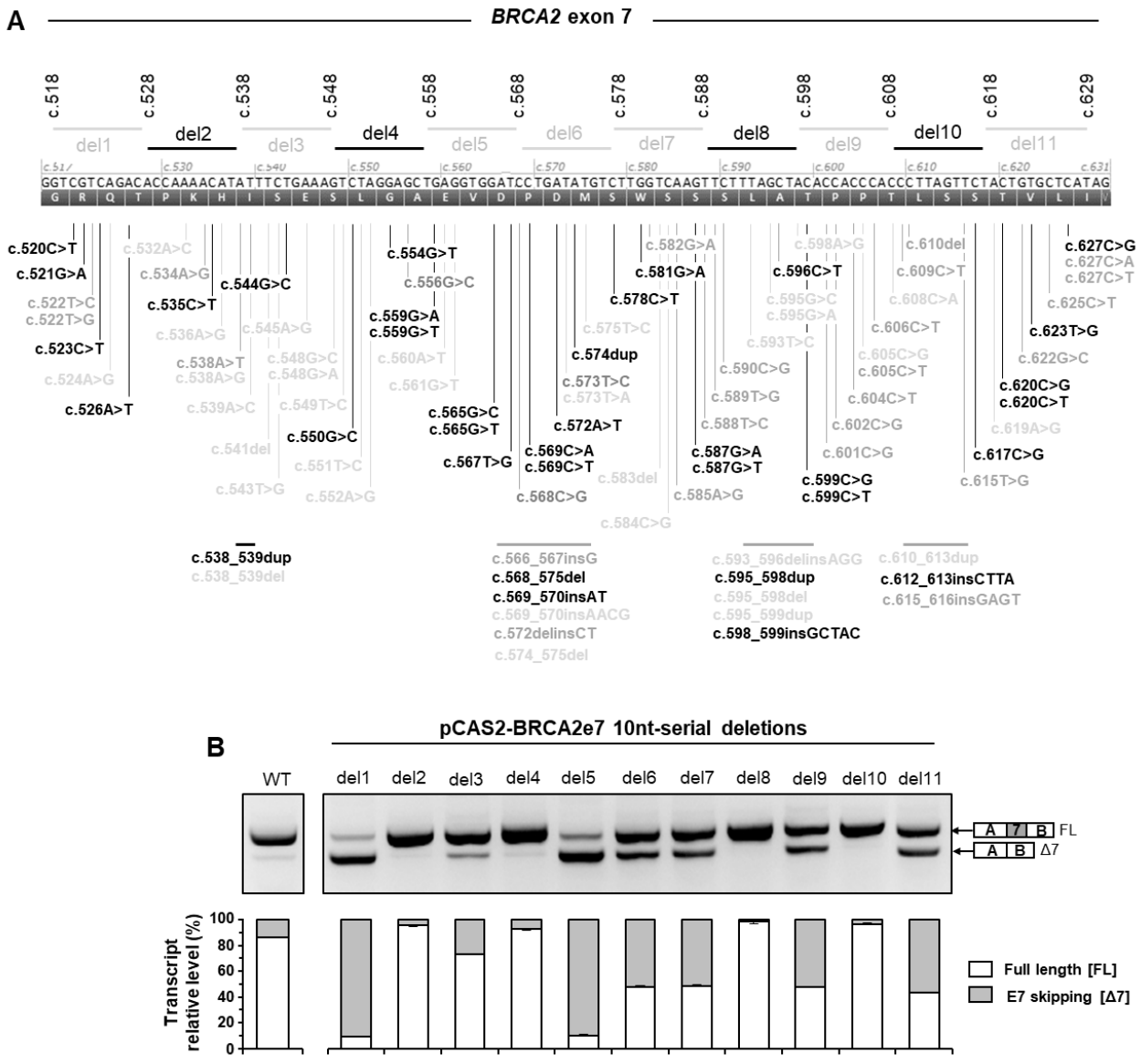


Figure 4. Mapping of short exonic regions by serial deletions revealed a relatively important density of splicing regulatory elements within *BRCA2* exon 7.

Variations			Splicing functional assay[1]						Reference[3]
Positions	Nucleotide variations[1]	Predicted protein changes	Minigene	Patient RNA analysis[2]					
				LCL$^{\alpha}$	LEU$^{\beta}$	STCL$^{\gamma}$	PBL^{c}	ASE	
	WT		85 ± 1.7	100 ± 0.0 99 ± 0.1	100 ± 0.3	100 ± 0.0 99 ± 0.1	100 ± 0.0 99 ± 0.1		
	c.517-1_631+462del		n/a	66 ± 1.0 55 ± 0.9	42 ± 0.5				
Intron 6 (n=4)	c.517-20A>G	p.?	24 ± 0.0	92 ± 0.3 77 ± 0.7	77 ± 0.3				
	c.517-19C>T	p.?	69 ± 2.2	99 ± 0.4 97 ± 0.8	96 ± 0.6				
	c.517-13_-9delTCTT	p.?	0 ± 0.0			56 ± 0.3 44 ± 0.4			
	c.517-1G>A	p.?	0 ± 0.0	68 ± 0.2 49 ± 0.1			65 ± 0.7 44 ± 0.9		
Exon 7 (n=14)	c.517G>T	p.Gly173Cys	1 ± 0.1	62 ± 0.9 42 ± 0.5				1 ± 0$^{\alpha}$	
	c.520C>T	p.Arg174Cys	10 ± 0.5	80 ± 0.2 60 ± 0.6	55 ± 0.0			41 ± 1 $^{\alpha}$ 24 ± 0 $^{\beta}$ 38 ± 1 $^{\beta}$	[1] [2] [3]
	c.521G>A	p.Arg174His	66 ± 2.9		nd			1 ± 0$^{\beta}$	
	c.538_539dup	p.Ile180Asnfs*6	69 ± 1.7	99 ± 0.0 97 ± 0.3					
	c.566A>G	p.Asp189Gly	0 ± 0.4	79 ± 0.0 49 ± 0.0				0 ± 0$^{\alpha}$	
	c.573T>C	p.=	88 ± 1.5		99 ± 0.0			97 ± 2$^{\beta}$	
	c.575T>C	p.Met192Thr	100 ± 0.0			100 ± 0.1 99 ± 0.1		107 ± 1$^{\gamma}$	
	c.587G>A	p.Ser196Asn	50 ± 1.2	97 ± 0.1 94 ± 0.5	91 ± 0.3			92 ± 1$^{\alpha}$ 69 ± 1$^{\beta}$	
	c.599C>T	p.Thr200Ile	41 ± 1.2	96 ± 0.2 84 ± 3.4			95 ± 0.5 85 ± 0.4	74 ± 2$^{\alpha}$	
	c.605C>G	p.Pro202Arg	96 ± 0.3			100 ± 0.1 99 ± 0.1		138 ± 1$^{\gamma}$	
	c.617C>G	p.Ser206Cys	25 ± 2.5		71 ± 0.0			49 ± 1$^{\beta}$	[1]
	c.625C>T	p.Leu209Phe	82 ± 0.7	100 ± 0.0 99 ± 0.1				178 ± 1$^{\alpha}$	
	c.627C>T	p.=	81 ± 0.3	99 ± 0.0 98 ± 0.0				99 ± 1$^{\alpha}$	
	c.631G>A	p.Val211Ile	0	75 ± 1.0 49 ± 0.8				4 ± 0$^{\alpha}$	
Intron 7 (n=4)	c.631+3A>G	p.?	0	76 ± 0.5 54 ± 1.0					
	c.631+25C>T	p.?	90 ± 1.0	99 ± 0.1 99 ± 0.1					
	c.631+29A>C	p.?	n/a	99 ± 0.1 99 ± 0.1					
	c.631+43G>T	p.?	n/a	100 ± 0.1 100 ± 0.0					

Table 1. Description of splicing outcomes of *BRCA2* exons 7 variants observed in patient's RNA samples.

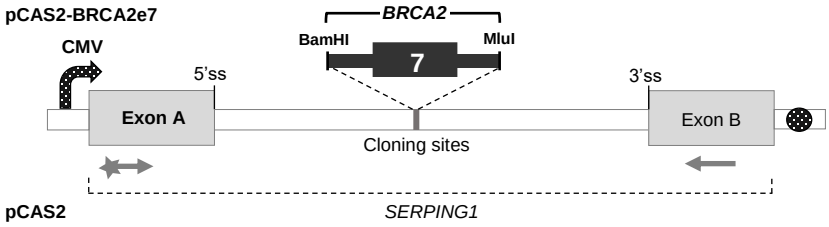
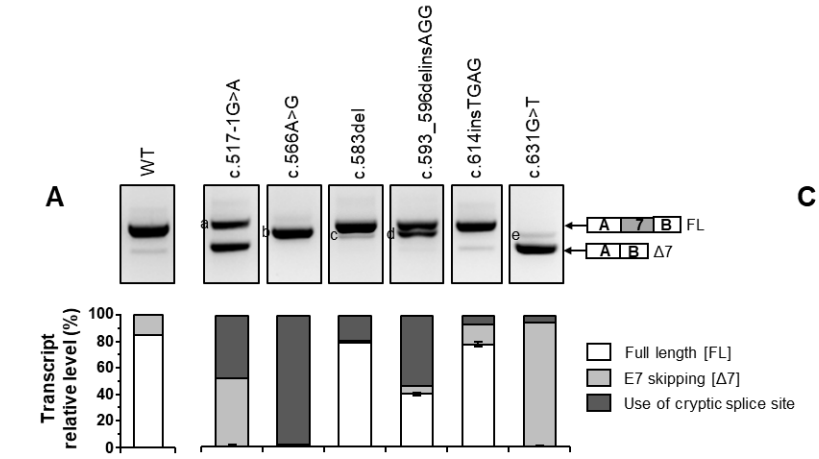


Figure S1. Structure of the pCAS2-*BRCA2*e7 minigene used in the splicing reporter assay.



C

Variation associated to the use of a cryptic splice site	Splicing defects due to the use of a cryptic splice site		*In silico* predictions	
			MES	SSFL
c.517-1G>A	a	Δp(1nt)	++	+
c.566A>G	b	Δq(70nt)	++	++
c.583del	c	Δq(50nt)	-	++
c.593_596delinsAGG	d	Δq(36nt)	+	-
c.614insTGAG		Δq(22nt)	+	+
c.631G>T	e	Δq(70nt)	-	-

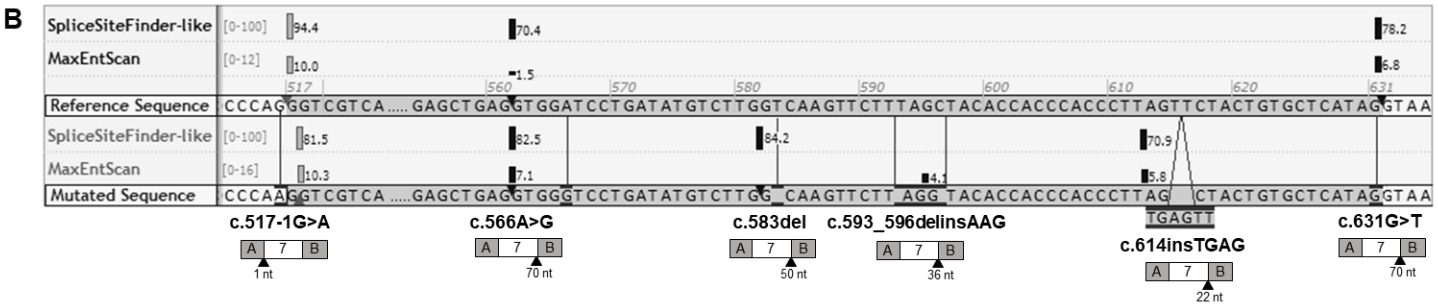


Figure S2. Splice-site dedicated bioinformatics predictions of variant-induced creation of de novo splice sites or activation of cryptic splice sites experimentally detected in the minigene assays.

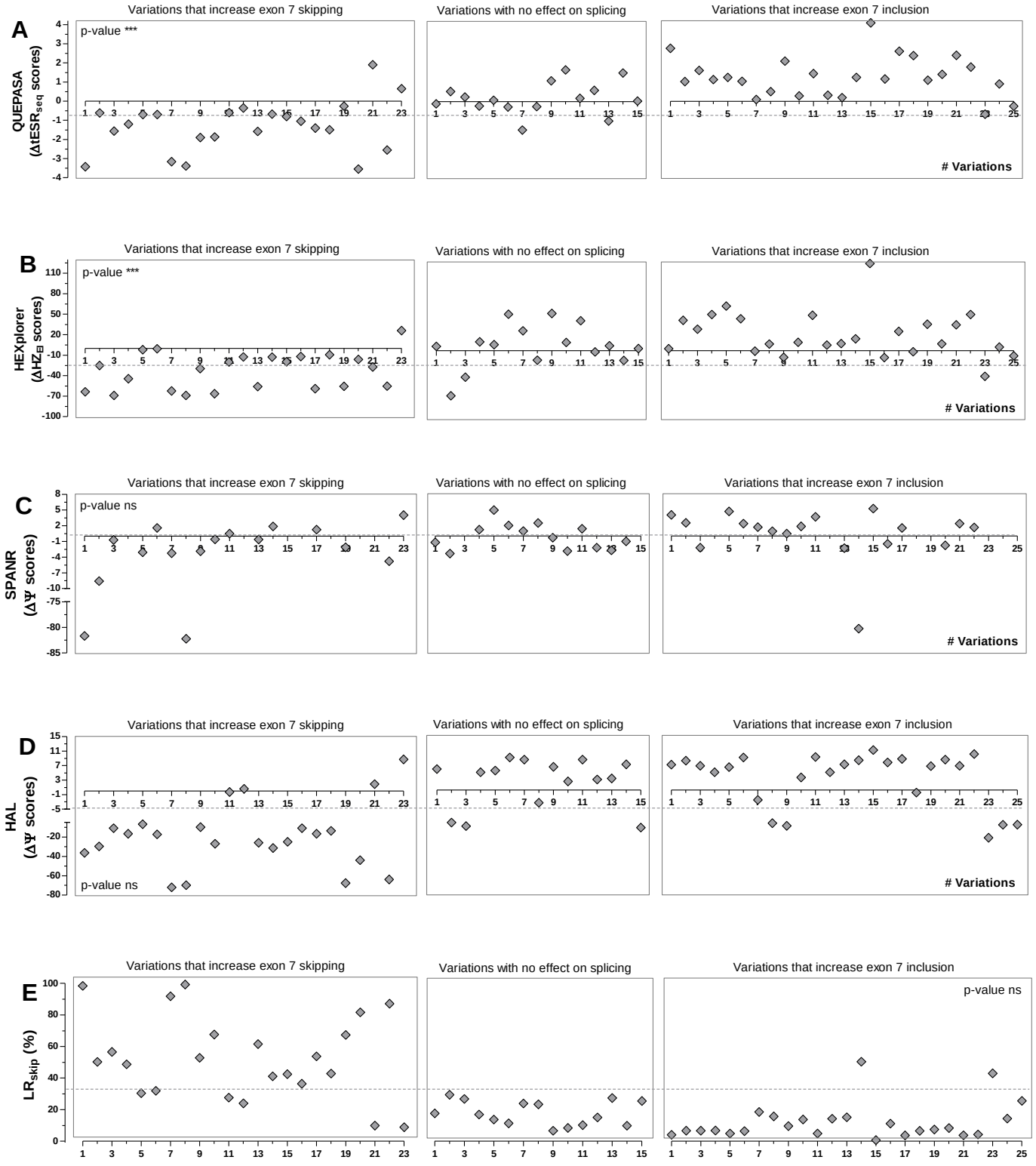


Figure S3. Comparison of exonic variant-associated splicing effects observed in the pCAS2-*BRCA2*e7 minigenes and associated SRE-dedicated *in silico* predictions.

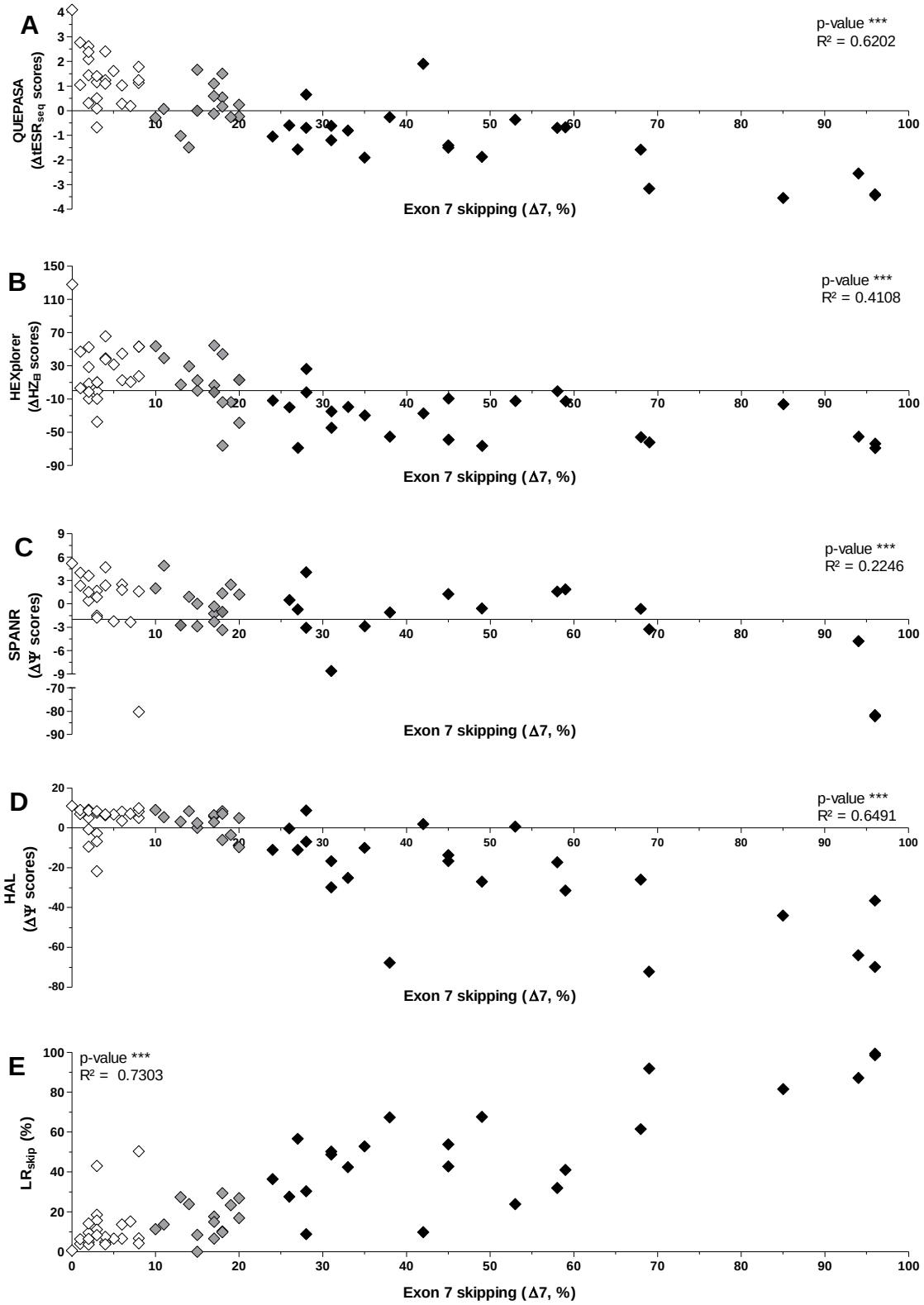


Figure S4. Correlation between variant-associated exon skipping levels described in the context of the pCAS2-BRCA2e7 minigene assays and *in silico* data obtained with SRE-dedicated approaches.

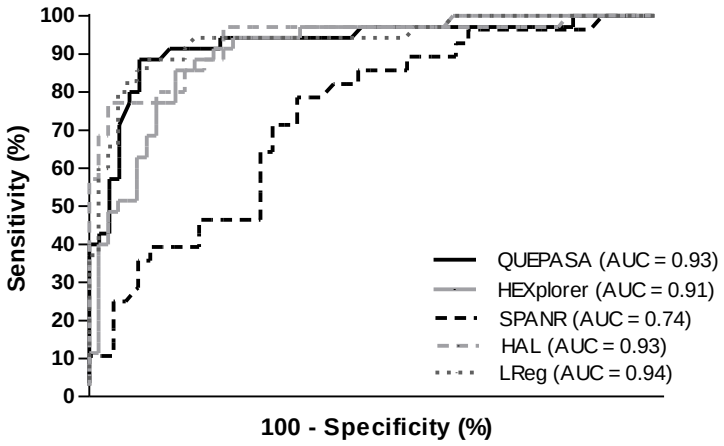


Figure S5. Receiver operating characteristic (ROC) curve of SRE-dedicated bioinformatics approach in predicting exon .

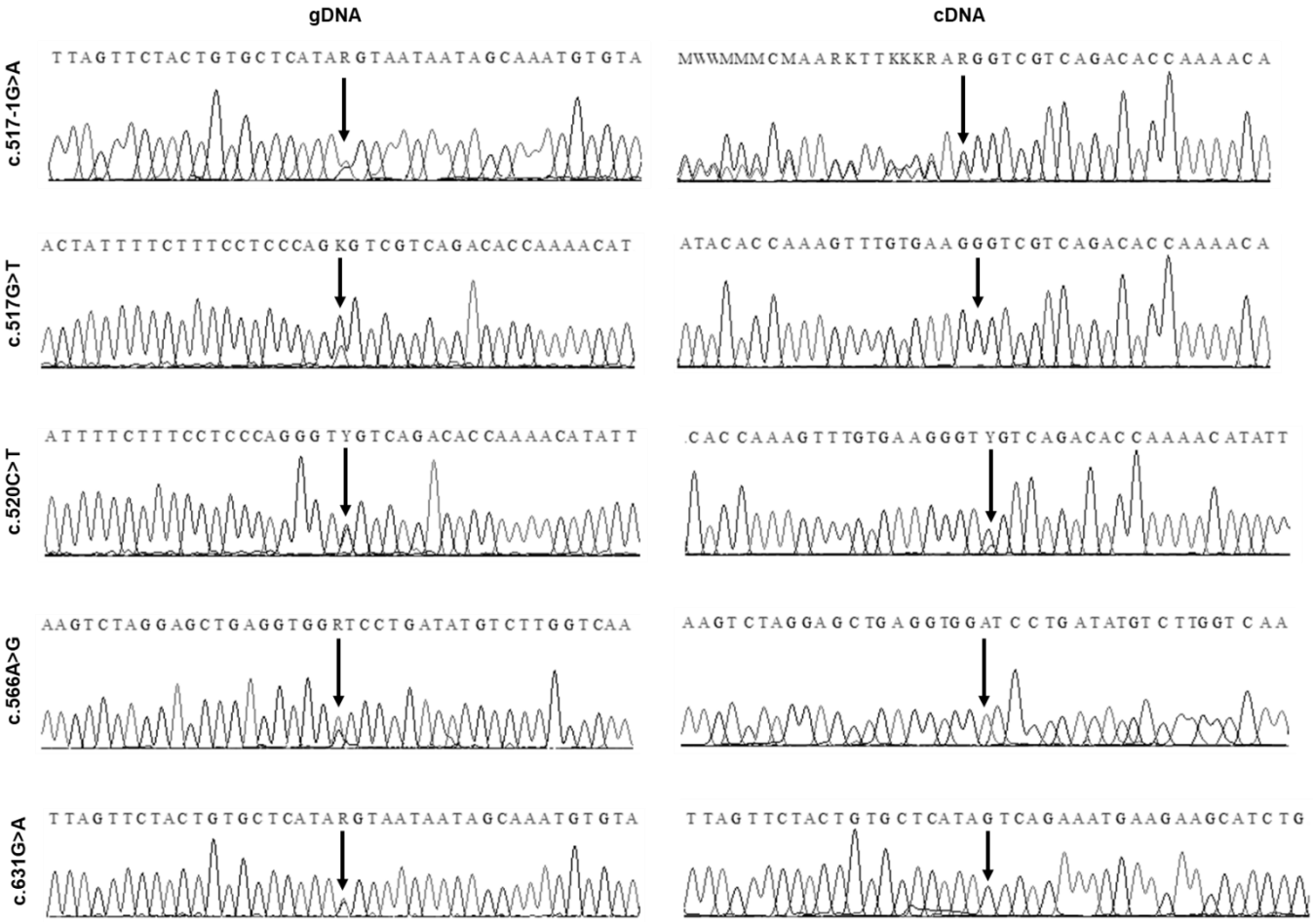


Figure S6. Sanger sequencing of the PCR and RT-PCR products of patients carrying variation causing near-total to total splicing defects.

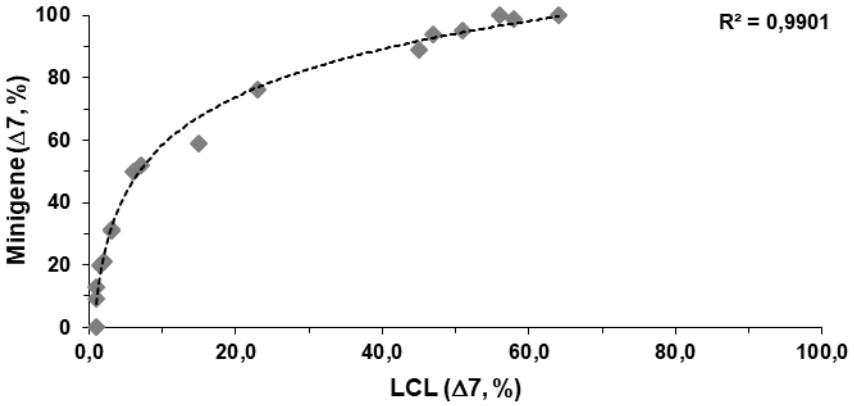


Figure S7. Correlation analysis between *BRCA2* exon 7 skipping levels observed in pCAS2-*BRCA2*e7 minigene assays and patient-derived lymphoblastoid cell lines.

Purpose		Name[1]	Sequence[2]
Splicing reporter minigene-based assay	Site-directed mutagenesis	Br2Ex7_c.517-7CT-F	CTATTTTCTTTCTTCCCAGGGTCGTC
		Br2Ex7_c.522TC-F	CTCCCAGGGTCGCCAGACACCAAAAC
		Br2Ex7_c.523CT-R	GTTTTGGTGTCTAACGACCCTGGGAG
		Br2Ex7_c.535CT-F	CAGACACCAAAATATATTTCTGAAAG
		Br2Ex7_c.536AG-F	GACACCAAAACGTATTTCTGAAAGTC
		Br2Ex7_c.538AT-R	GACTTTCAGAAAAATGTTTTGGTGTC
		Br2Ex7_c.538AG-R	GACTTTCAGAAACATGTTTTGGTGTC
		Br2Ex7_c.543TG-F	CCAAAACATATTTCGGAAAGTCTAGGAGC
		Br2Ex7_c.544GC-R	CTCCTAGACTTTGAGAAATATG
		Br2Ex7_c.545AG-F	CATATTTCTGGAAGTCTAGGAGCTG
		Br2Ex7_c.548GA-R	CAGCTCCTAGATTTTCAGAAATATG
		Br2Ex7_c.550CG-F	CATATTTCTGAAAGTGTAGGAGCTGAGG
		Br2Ex7_c.552AG-R	CCACCTCAGCTCCCAGACTTTCAG
		Br2Ex7_c.554GT-F	CTGAAAGTCTAGTAGCTGAGGTGGATC
		Br2Ex7_c.556GC-R	GATCCACCTCAGGTCCTAGACTTTC
		Br2Ex7_c.559GT-R	CAGGATCCACCTAAGCTCCTAGACTTTC
		Br2Ex7_c.560AT-R	TCAGGATCCACCACAGCTCCTAGACTTTC
		Br2Ex7_c.561GT-F	CTAGGAGCTGATGTGGATCCTGATATG
		Br2Ex7_c.565GC-R	GACATATCAGGATGCACCTCAGCTCC
		Br2Ex7_c.566_567insG-F	AGCTGAGGTGGAGTCCTGATATGTC
		Br2Ex7_c.566AG-F	GAGCTGAGGTGGGTCCTGATATGTC
		Br2Ex7_c.568_575del-F	GCTGAGGTGGATGTCTTGGTCAAG
		Br2Ex7_c.568CG-R	CAAGACATATCAGCATCCACCTCAG
		Br2Ex7_c.569_570insAACG-F	GAGGTGGATCCAACGTGATATGTC
		Br2Ex7_c.569CA-R	CAAGACATATCATGATCCACCTCAG
		Br2Ex7_c.569CT-F	CTGAGGTGGATCTTGATATGTCTTG
		Br2Ex7_c.574dup-F	GTGGATCCTGATAATGTCTTGGTCAAG
		Br2Ex7_c.578CT-F	GATCCTGATATGTTTTGGTCAAGTTC
		Br2Ex7_c.583delT-F	GATATGTCTTGGCAAGTTCTTTAG
		Br2Ex7_c.584CG-R	GCTAAAGAACTTCACCAAGACATATC
		Br2Ex7_c.585AG-F	GTCTTGGTCGAGTTCTTTAGCTAC
		Br2Ex7_c.589TG-R	GTGTAGCTAAAGCACTTGACCAAGAC
		Br2Ex7_c.590CG	GTGTAGCTAAACAACTTGACCAAGAC
		Br2Ex7_c.593_596delinsAGG-R	GGGTGGTGTACCTTAAGAACTTGACC
		Br2Ex7_c.593TC-F	GGTCAAGTTCTTCAGCTACACCACC
		Br2Ex7_c.594_598dup-F	AGTTCTTTAGCTAAGCTACACCACCC
		Br2Ex7_c.595_598del-R	GGGTGGGTGGTGTAAAGAACTTGAC
		Br2Ex7_c.595_599dup-R	GTGGGTGGTGTAGCGTAGCTAAAG
		Br2Ex7_c.595GA-F	CAAGTTCTTTAACTACACCACCCAC
		Br2Ex7_c.595GC-R	GTGGGTGGTGTAGGTAAAGAACTTG
		Br2Ex7_c.598_599insGCTAC-F	GTTCTTTAGCTAGCTACCACCACCCAC
		Br2Ex7_c.601CG-R	CTAAGGGTGGGTGCTGTAGCTAAAG
		Br2Ex7_c.605CT-R	GAACTAAGGGTGAGTGGTGTAGC
		Br2Ex7_c.606CT-F	GCTACACCACCTACCCTTAGTTCTAC
		Br2Ex7_c.608CA-R	GTAGAACTAAGGTTGGGTGGTGTAGC
		Br2Ex7_c.610_613dup-F	CCACCCACCCTTACTTAGTTCTACTG
		Br2Ex7_c.612_613insCTTA-R	CAGTAGAACTTAAGAAGGGTGGGTG
		Br2Ex7_c.615TG-R	GAGCACAGTAGACCTAAGGGTGGG
		Br2Ex7_c.631GT-R	GCTATTATTACATATGAGCACAGTAG
	Serial deletions	Br2Ex7-del-1-F	CTATTTTCTTTCCTCCCAGGGACCAAAACATATTTCTGAAAGT

		Br2Ex7-del-2-F	CCTCCCAGGGTCGTCAGACATTTCTGAAAGTCTAGGAGC
		Br2Ex7-del-3-F	CGTCAGACACCAAAACATGTCTAGGAGCTGAGGTGGATC
		Br2Ex7-del-4-F	CCAAAACATATTTCTGAAATGAGGTGGATCCTGATATGT
		Br2Ex7-del-5-F	TTTCTGAAAGTCTAGGAGCCCTGATATGTCTTGGTCAAG
		Br2Ex7-del-6-F	GGAGCTGAGGTGGATCTTGGTCAAGTTCTTTAGCTAC
		Br2Ex7-del-7-R	GGTGGTGTAGCTAAAGAAACATATCAGGATCCACCTC
		Br2Ex7-del-8-R	CTAAGGGTGGGTGGTGTCTTGACCAAGACATATCAGG
		Br2Ex7-del-9-R	GCACAGTAGAACTAAGGGAGCTAAAGAACTTGACCAAG
		Br2Ex7-del-10-R	TTACCTATGAGCACAGTATGGGTGGTGTAGCTAAAGAAC
		Br2Ex7-del-11-R	ATTTGCTATTATTACCTAGAACTAAGGGTGGGTGGTG
	PCR (cloning, minigene preparation)	Br2Ex7_InFus_BamHI-F	AGGCTAAGAAGTGCAGGATCTGTTATACCTTTGCCCTGA GATTTAC
		Br2Ex7_InFus_MluI-R	AGGGGTCAAAACAAGACGCGTTGCTTGACACCACTGGAC TA
	Sequencing of minigene inserts	pCAS-Seq-F	GGGGTCAATAGCAGTGAGAG
		pCAS-Seq-R	GCTCCATTTCACAGGTAGAGA
	RT-PCR and sequencing of RT-PCR products	6FAM-pCAS-KO1-F (5'-fluo)	TGACGTCGCCGCCCATCAC
		pCAS-2R	ATTGGTTGTTGAGTTGGTTGTC
Allele specific expression (SNaPshot®)	Primer extension	Br2Ex7_c.517GT-R	GAAATATGTTTTGGTGTCTGACGAC
		Br2Ex7_c.520CT-R	TTTCAGAAATATGTTTTGGTGTCTGAC
		Br2Ex7_c.521GA-R	CTTTCAGAAATATGTTTTGGTGTCTGA
		Br2Ex7_c.538_539dup-F	GGTCGTCAGACACCAAAACATAT
		Br2Ex7_c.538_539dup-R	CCTCAGCTCCTAGACTTTCAGAAA
		Br2Ex7_c.566AG-F	GAAAGTCTAGGAGCTGAGGTGG
		Br2Ex7_c.566AG-R	GAACTTGACCAAGACATATCAGGA
		Br2Ex7_c.573TC-F	GGAGCTGAGGTGGATCCTGA
		Br2Ex7_c.573TC-R	GTAGCTAAAGAACTTGACCAAGACAT
		Br2Ex7_c.575TC-F	GGAGCTGAGGTGGATCCTGATA
		Br2Ex7_c.575TC-R	TGTAGCTAAAGAACTTGACCAAGAC
		Br2Ex7_c.587GA-F	GTGGATCCTGATATGTCTTGGTCAA
		Br2Ex7_c.587GA-R	GGTGGGTGGTGTAGCTAAAGAA
		Br2Ex7_c.599CT-F	GATATGTCTTGGTCAAGTTCTTTAGCTA
		Br2Ex7_c.599CT-R	AGTAGAACTAAGGGTGGGTGGT
		Br2Ex7_c.605TC-F	GGTCAAGTTCTTTAGCTACACCAC
		Br2Ex7_c.605TC-R	TATGAGCACAGTAGAACTAAGGGTG
		Br2Ex7_c.617TC-F	CTACACCACCCACCCTTAGTT
		Br2Ex7_c.625CT-F	ACCCACCCTTAGTTCTACTGTG
		Br2Ex7_c.627CT-F	CCACCCTTAGTTCTACTGTGCT
		Br2Ex7_c.631GA-F	ACCCTTAGTTCTACTGTGCTCATA
	RT-PCR on patient RNA	Br2Ex6_F	AGTGGTATGTGGGAGTTTGTTTC
		Br2Ex9_R	TTTCTTCAGACTTTCATCATGATTGG
	PCR on genomic DNA	Br2Ex7_InFus_BamHI-F	AGGCTAAGAAGTGCAGGATCTGTTATACCTTTGCCCTGA GATTTAC
		Br2Ex7_InFus_MluI-R	AGGGGTCAAAACAAGACGCGTTGCTTGACACCACTGGAC TA
Patient RNA analysis	RT-PCR on patient RNA	Br2Ex6_F	AGTGGTATGTGGGAGTTTGTTTC
		6FAM-Br2Ex9_R (5'-fluo)	TTTCTTCAGACTTTCATCATGATTGG

Table S1. Description of the primers used in this study.

Variation	Cells / tissue (collection method)	Culture conditions	Nonsense mediated decay inhibition	RNA extraction method	Dnase 1 treatment	Amount of total RNA used in cDNA synthesis	cDNA synthesis primer	cDNA synthesis protocol	PCR amplification template	PCR conditions[2]	PCR product analysis	Sequencing
c.517-1_631+462del	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mM L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.517-20A>G	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel) [1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.517-19C>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500

	cDNA (A. Solano, Argentina)											
c.517-13_-9delTCTT	Short-term cultured lymphocytes (A. Vega, Spain)	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a			
c.517-1G>A	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
	PBL (C. Lazaro and M. Menendez, Spain)											
c.517G>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.520C>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500

										then 72ºC, 10 min		
	cDNA (D. Baralle, UK)											
	(B. Wappenschmit)											
c.521G>A	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.538_539dup	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.566A>G	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.573T>C	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.575T>C	Short-term cultured lymphocytes with phytohaemaglut					200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1]	Sequencing of cDNA purified bands using ABI3500

	inin (R. Blok, The Netherlands)									94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Capillary electrophoresis using ABI3500	
c.587G>A	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a		Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.599C>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
	PBL (C. Lazaro and M. Menendez, Spain)											
c.605C>G	Short-term cultured lymphocytes with phytohaemaglutinin (R. Blok, The Netherlands)					200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.617C>G	Fresh whole peripheral blood	n/a	n/a	PAXgene Blood RNA kit (Qiagen)[1]	RNase-Free DNase I (Qiagen)[1]	200ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 63-50ºC, 30s 72ºC, 1 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1]	Sequencing of cDNA purified bands using ABI3500

										then 26 cycles of 94ºC, 30s 50ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Capillary electrophoresis using ABI3500	
c.625C>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.627C>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.631G>A	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.631+3A>G	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.631+25C>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500
c.631+29A>C	EBV-immortalized	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	2.5x10^6 cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1]	Sequencing of cDNA purified bands using ABI3500

	lymphoblastoid cell line									then 72ºC, 10 min	Capillary electrophoresis using ABI3500	
c.631+43G>T	EBV-immortalized lymphoblastoid cell line	RPMI + 10% fetal calf serum + 2 mMof L-glutamine	$2.5x10^6$ cells 200 µg /ml 5.5 hours	NucleoSpin RNA II kit (Macherey Nagel)[1]	RNase-Free DNase I (Marchery Nagel)[1]	200 ng	Gene specific	One step RT-PCR (Qiagen) 50°C for 30 min	n/a	95ºC, 15 min then 26 cycles of 94ºC, 30s 55ºC, 30s 72ºC, 1 min then 72ºC, 10 min	Agarose gel and gel purification of aberrant bands using NucleoSpin Gel and PCR Clean-up (Macherey Nagel)[1] Capillary electrophoresis using ABI3500	Sequencing of cDNA purified bands using ABI3500

Table S2. Description of protocols used to analyse patient RNA samples.

Purpose	Gaussian distribution (Shapiro Wilk test)	Statistical analysis	Bioinformatics approach
Linear correlation between exon inclusion levels and *in silico* predictions	Yes	Spearman	QUEPASA, HEXplorer
	No	Pearson	SPANR, HAL, LR_{skip}, LR_{inc}
Discrimination of 3 groups of variants (↑ exon skipping *versus* no effect on splicing versus ↑ exon inclusion).	Yes	Anova (Bonferroni post-tests)	QUEPASA, HEXplorer
	No	Kruskal-Wallis (Duns post-tests)	SPANR, HAL, LR_{skip}, LR_{inc}
Discrimination of 2 groups of variants (variants that increase exon skipping or exon inclusion versus those that do not)	Yes	Student	QUEPASA, HEXplorer
	Yes	Student with Welsh's correction	
	No	Mann-Whitney	SPANR, HAL, LR_{skip}, LR_{inc}

Table S3. Description of statistical analyses conducted in this study.

			Databases Clinical classification[1]										Splicing functional assay[2]						Splicing data		
Positions	Nucleotide variations[1]	Predicted protein changes	ENIGMA	BIC	BRCA-Share	COSMIC	ClinVar	dbSNP	ESP	gnmoAD	HGMD	LOVD	Minigene	Patient RNA analysis: LCL $^{\alpha}$	LEU $^{\beta}$	STCL $^{\gamma}$	PBL ^{c}	ASE	Effect on splicing[3]	Classification suggested	References[4]
	WT												85 ± 1.7	100 ± 0.0 99 ± 0.1	100 ± 0.3	100 ± 0.0 99 ± 0.1	100 ± 0.0 99 ± 0.1		-	-	
	c.517-1_631+462del				5								n/a	66 ± 1.0 55 ± 0.9	42 ± 0.5				Total effect (Δ7)	5	
Intron 6 (n=8)	c.517-22_-23del	p.?	n/a		2								60 ± 0.7						Partial effect (Δ7)	3	
	c.517-20A>G	p.?	n/a		3		2					n/a	24 ± 0.0	92 ± 0.3 77 ± 0.7	77 ± 0.3				Partial effect (Δ7)	3	
	c.517-19C>T	p.?	1	n/a	1		1	1	n/a	n/a			69 ± 2.2	99 ± 0.4 97 ± 0.8	96 ± 0.6				Partial effect (Δ7)	3	
	c.517-13_-9delTCTT	p.?	n/a										0 ± 0.0			56 ± 0.3 44 ± 0.4			Total effect (Δ7)	5	
	c.517-7C>T	p.?	n/a										98 ± 0.3						Partial effect (+E7)	1	
	c.517-4C>G	p.?	n/a	n/a			1 3	1 2 3	3	n/a		n/a	98 ± 0.3						Partial effect (+E7)	1	
	c.517-2A>G	p.?	n/a	n/a	5		4 5	3 4 5			5		0 ± 0.0						Total effect (Δ7)	5	
	c.517-1G>A	p.?	n/a	n/a	5		5	5			5		0 ± 0.0	68 ± 0.2 49 ± 0.1			65 ± 0.7 44 ± 0.9		Total effect (Δ7 = 52%-7% ; Δ7p(1nt) = 48-44% ; in minigene and patient, respectively)	5	
Exon 7 (n=)	c.517G>T*	p.Gly173Cys			3		3	3			5	n/a	1 ± 0.1	62 ± 0.9 42 ± 0.5				1 ± 0 $^{\alpha}$	Total effect (Δ7)	5	[1] [2]
	c.517G>C	p.Gly173Arg	n/a				3 4 5	3			5		4 ± 0.4						Total effect (Δ7)	5	
	c.518G>T*	p.Gly173Val	n/a	n/a	3		3	3		n/a	5	n/a	98 ± 0.4						Partial effect (+E7)	3	[2]
	c.518del*	p.Gly173Valfs*12		5	5		5	5	n/a	n/a	5		93 ± 0.1						Partial effect (+E7)	5	[2]
	c.520C>T*	p.Arg174Cys	n/a		2	n/a	3	3		n/a	5	n/a	10 ± 0.5	80 ± 0.2 60 ± 0.6	55 ± 0.0			41 ± 1 $^{\alpha}$ 24 ± 0 $^{\beta}$ 38 ± 1 $^{\beta}$	Partial effect (Δ7)	3	[1][2][4]
	c.521G>A*	p.Arg174His	n/a	n/a	3		2 3	3		n/a	3		66 ± 2.9		nd			1 ± 0 $^{\beta}$	Partial effect (Δ7)	3	[2]
	c.522T>C	p.=				n/a							83 ± 0.7						No effect	1	
	c.522T>G*	p.=	n/a		3							n/a	86 ± 1.8						No effect	1	[1] [2]
	c.523C>T	p.Gln175*					5	5					4 ± 0.3						Total effect (Δ7)	5	
	c.524A>G	p.Gln175Arg			3								99 ± 0.1						Total effect (+E7)	3	
	c.526A>T	p.Thr176Ser	n/a										69 ± 3.6						Partial effect (Δ7)	3	
	c.532A>C*	p.Lys178Gln			3								97 ± 0.6						Partial effect (+E7)	3	[2]
	c.534A>G*	p.=	1		3		1 2	1 2		n/a			83 ± 1.4						No effect	1	[2]
	c.535C>T	p.His179Tyr											73 ± 2.3						Partial effect (Δ7)	3	
	c.536A>G	p.His179Arg	n/a					n/a		n/a			94 ± 0.9						Partial effect (+E7)	3	

	c.538A>T	p.Ile180Phe			3		3						82 ± 2.0						No effect	3	
	c.538A>G	p.Ile180Val					3						95 ± 0.5						Partial effect (+E7)	3	
	c.538_539dup	p.Ile180Asnfs*6	5	5			5	5			5		69 ± 1.7	99 ± 0.0 97 ± 0.3					Partial effect (Δ7)	5	
	c.538_539del*	p.Ile180Phefs*2		5	5		5	5			5		93 ± 0.8						Partial effect (+E7)	5	[2]
	c.539T>C*	p.Ile180Thr	n/a	n/a			3	3		n/a			97 ± 0.5						Partial effect (+E7)	3	[2]
	c.541del	p.Ser181Leufs*4	5				5						93 ± 1.1						Partial effect (+E7)	5	
	c.[543T>G+560A>T]	p.Glu187Val	n/a										99 ±0.1						Total effect (+E7)	3	
	c.543T>G	p.=											97 ± 0.3						Partial effect (+E7)	1	
	c.544G>C	p.Glu182Gln					3	3					72 ± 4.4						Partial effect (Δ7)	3	
	c.545A>G	p.Glu182Gly					3	3					99 ± 0.0						Total effect (+E7)	3	
	c.548G>A	p.Ser183Asn					3	3					97 ± 0.7						Partial effect (+E7)	3	
	c.548G>C	p.Ser183Thr	n/a										97 ± 0.4						Partial effect (+E7)	3	
	c.549T>C*	p.=		n/a			2	3					94 ± 0.8						Partial effect (+E7)	1	[2]
	c.550C>G	p.Leu184Val					3	3		n/a			42 ± 1.4						Partial effect (Δ7)	3	
	c.551T>C*	p.Leu184Pro		n/a			3	2					99 ± 0.3						Total effect (+E7)	3	[2]
	c.552A>G	p.=				n/a							99 ± 0.3						Total effect (+E7)	1	
	c.554G>T	p.Gly185Val					3	3					31 ± 1.5						Partial effect (Δ7)	3	
	c.556G>C*	p.Ala186Pro			3		3						84 ± 4.5						No effect	3	[3]
	c.559G>T	p.Glu187*				n/a							4 ± 0.2						Total effect (Δ7)	5	
	c.559G>A*	p.Glu187Lys		n/a		n/a	3	3			3		37 ± 1.9						Partial effect (Δ7)	3	[2]
	c.560A>T	p.Glu187Val											94 ± 0.4						Partial effect (+E7)	3	
	c.561G>T	p.Glu187Asp						n/a		n/a			98 ± 0.4						Partial effect (+E7)	3	
	c.565G>C	p.Asp189His				n/a							65 ± 3.1						Partial effect (Δ7)	3	
	c.565G>T	p.Asp189Tyr	n/a				3						51 ± 3.2						Partial effect (Δ7)	3	
	c.566A>G	p.Asp189Gly			3		3	3					0 ± 0.4	79 ± 0.0 49 ± 0.0				$0 \pm 0^{\alpha}$	Total effect (Δ7q(70nt))	5	
	c.566_567insG	p.Asp189Glufs*3					5						80 ± 0.3						No effect	5	
	c.567T>G	p.Asp189Glu	n/a										75 ± 2.4						Partial effect (Δ7)	3	
	c.568C>G	p.Pro190Ala					3	3					80 ± 0.9						No effect	3	
	c.568_575del	p.Pro190Valfs*13					5	5					47 ± 0.7						Partial effect (Δ7)	5	
	c.569C>T	p.Pro190Leu					3	3			3		32 ± 1.2						Partial effect (Δ7)	3	
	c.569C>A	p.Pro190His				n/a							41 ± 1.8						Partial effect (Δ7)	3	
	c.569_570insAACG	p.Asp191Thrfs*2					5	5					98 ± 0.0						Partial effect (+E7)	5	
	c.569_570insAT	p.Asp191Leufs*9	5										67 ± 0.7						Partial effect (Δ7)	5	
	c.572A>G*	p.Asp191Gly			3		n/a	3			5	n/a	5 ± 0.0						Total effect (Δ7q(60nt))	3	[1] [2]
	c.572A>T*	p.Asp191Val	n/a		3		3	3		n/a	5	n/a	72 ± 1.7						Partial effect (Δ7)	3	[2]
	c.572delinsCT*	p.Asp191Alafs*15	5		5		5				5		84 ± 1.4						No effect	5	[2]
	c.573T>A	p.Asp191Glu					3				3		93 ± 1.0						Partial effect (+E7)	3	
	c.573T>C*	p.=			2		2			n/a			88 ± 1.5		99 ± 0.0			$97 \pm 2^{\beta}$	No effect	1	[2]
	c.574dup	p.Met192Asnfs*14					5	5			5		76 ± 1.9						Partial effect (Δ7)	5	

	c.574_575del*	p.Met192Valfs*13	5	5			5				5		96 ± 0.6						Partial effect (+E7)	5	[2]
	c.575T>C*	p.Met192Thr	n/a	n/a		n/a	3	2 3			3		100 ± 0.0			100 ± 0.1 99 ± 0.1		107 ± 1$^{\gamma}$	Total effect (+E7)	3	[2]
	c.578C>T	p.Ser193Phe					3	3					55 ± 2.4						Partial effect (Δ7)	3	
	c.581G>A*	p.Trp194*		5			5	5			5	n/a	13 ± 1.8						Partial effect (Δ7)	5	[2]
	c.582G>A*	p.Trp194*		5	5		5	5			5		84 ± 1.3						No effect	5	[2]
	c.583del	p.Ser195Glnfs*4					5	5			5		79 ± 0.0						Partial effect (Δ7q(50nt) = 19% ; Δ7 = 2%)	5	
	c.584C>G	p.Ser195*	5				5	5			5		92 ± 0.5						Partial effect (+E7)	5	
	c.585A>G	p.=			3								89 ± 0.8						No effect	1	
	c.587G>T*	p.Ser196Ile		n/a		n/a	3	3		n/a	3		48 ± 1.9						Partial effect (Δ7)	3	[2]
	c.587G>A*	p.Ser196Asn		n/a	3	n/a	3	3			5	n/a	50 ± 1.2	97 ± 0.1 94 ± 0.5	91 ± 0.3			92 ± 1$^{\alpha}$ 69 ± 1$^{\beta}$	Partial effect (Δ7)	3	[1][2]
	c.588T>C	p.=			3		2	2					90 ± 1.4						No effect	1	
	c.589T>G	p.Ser197Ala					3	3					86 ± 0.5						No effect	3	
	c.590C>G	p.Ser197Cys					3	3					81 ± 1.2						No effect	3	
	c.593T>C	p.Leu198Ser						n/a		n/a			100 ± 0.1						Total effect (+E7)	3	
	c.593_596delinsAGG	p.Leu198*					5				5		40 ± 1.0						Partial effect (Δ7q(36nt), 54% Δ7 = 6%)	5	
	c.594_598dup	p.Thr200Lysfs*13					5	5					55 ± 1.4						Partial effect (Δ7)	5	
	c.595G>A	p.Ala199Thr					3	3	n/a				97 ± 0.6						Partial effect (+E7)	3	
	c.595C>G	p.Ala199Pro					3	3		n/a			98 ± 0.1						Partial effect (+E7)	3	
	c.595_598del	p.Ala199Hisfs*11					5	5					98 ± 0.1						Partial effect (+E7)	5	
	c.595_599dup	p.Pro201Leufs*12					5				5		96 ± 0.7						Partial effect (+E7)	5	
	c.596C>T	p.Ala199Val	n/a				3	3					62 ± 2.5						Partial effect (Δ7)	3	
	c.598A>G	p.Thr200Ala	n/a				3	3					97 ± 0.5						Partial effect (+E7)	3	
	c.599C>A*	p.Thr200Lys			3						3		46 ± 2.4						Partial effect (Δ7)	3	[2]
	c.599C>T*	p.Thr200Ile	n/a		3		3	3		n/a	3		41 ± 1.2	96 ± 0.2 84 ± 3.4			95 ± 0.5 85 ± 0.4	74 ± 2$^{\alpha}$	Partial effect (Δ7)	3	[2]
	c.598_599insGCTAC	p.Thr200Serfs*13						5					15 ± 0.8						Partial effect (Δ7)	5	
	c.601C>G*	p.Pro201Ala											87 ± 0.7						No effect	3	[3]
	c.602C>G*	p.Pro201Arg					n/a	n/a			5		84 ± 1.6						No effect	3	[2]
	c.604C>T	p.Pro202Ser					3	3					83 ± 1.7						No effect	3	
	c.605C>G	p.Pro202Arg	n/a										96 ± 0.3			100 ± 0.1 99 ± 0.1		138 ± 1$^{\gamma}$	Partial effect (+E7)	3	
	c.605C>T	p.Pro202Leu			3	n/a	3						85 ± 0.9						No effect	3	
	c.606C>T	p=					3	3		n/a			82 ± 2.3						No effect	1	
	c.608C>A	p.Thr203Asn					3	3		n/a			92 ± 1.0						Partial effect (+E7)	3	
	c.609C>T	p.=	n/a				2			n/a			83 ± 1.9						No effect	1	
	c.610delC*	p.Ser205Valfs*6		5			5	5			5		82 ± 0.1						No effect	5	[2]
	c.610_613dup	p.Ser205Thrfs*2					5	5			5		97 ± 0.6						Partial effect (+E7)	5	
	c.612_613insCTTA	p.Ser205Leufs*11											58 ± 2.2						Partial effect (Δ7)	5	

	c.615_616insGAGT	p.Ser206Glufs*10			5		5						78 ± 1.8						Partial effect (Δ7q(22nt) = 7% ; Δ7 = 15%)	5	
	c.615T>G	p.Ser205Arg											87 ± 1.5						No effect	3	
	c.617C>G*	p.Ser206Cys			3		n/a	n/a			5	n/a	25 ± 2.5		71 ± 0.0			$49 \pm 1^{\beta}$	Partial effect (Δ7)	3	[1][2]
	c.619A>G*	p.Thr207Ala	n/a	n/a			3	3		n/a	3	n/a	93 ± 0.5						Partial effect (+E7)	3	[2]
	c.620C>G	p.Thr207Ser	n/a										5 ± 0.3						Total effect (Δ7)	5	
	c.620C>T*	p.Thr207Ile	n/a				3	3		n/a	3		21 ± 0.3						Partial effect (Δ7)	3	[2]
	c.622G>C	p.Val208Leu	n/a										82 ± 1.5						No effect	3	
	c.623T>G*	p.Val208Gly		n/a			3	3		n/a	3		75 ± 1.3						Partial effect (Δ7)	3	[2]
	c.625C>T*	p.Leu209Phe			3							n/a	82 ± 0.7	100 ± 0.0 99 ± 0.1				$178 \pm 1^{\alpha}$	No effect	3	[1][2]
	c.627C>T*	p.=	2		2		1 2	1 2		n/a			81 ± 0.3	99 ± 0.0 98 ± 0.0				$99 \pm 1^{\alpha}$	No effect	1	[2]
	c.627C>A*	p.=	2		2		1 2	1 2		n/a			87 ± 1.4						No effect	1	[2]
	c.627C>G	p.=			2								72 ± 2.2						Partial effect (Δ7)	3	
	c.631G>A*	p.Val211Ile	n/a	n/a	5		5	5			5	n/a	0	75 ± 1.0 49 ± 0.8				$4 \pm 0^{\alpha}$	Total effect (Δ7q(22nt) = 5% ; Δ7 = 95%)	5	[1][2]
	c.631G>C*	p.Val211Leu		n/a			4 5	5			3		0						Total effect (Δ7q(22nt) = 5% ; Δ7 = 95%)	5	[2]
	c.631G>T	p.Val211Phe	n/a										0						Total effect (Δ7q(22nt) = 5% ; Δ7 = 95%)	5	
Intron 7 (n=7)	c.631+2T>G	p.?	n/a	n/a			5	5			5	n/a	0						Total effect (Δ7q(22nt) = 4% ; Δ7 = 96%)	5	
	c.631+3A>G	p.?	n/a		5		5	3 5			3		0	76 ± 0.5 54 ± 1.0					Total effect (Δ7q(22nt) = 5% ; Δ7 = 95%)	5	[5]
	c.631+7A>G	p.?	n/a				2 3	2 3					68 ± 1.4						Partial effect (Δ7)	3	
	c.631+25C>T	p.?	n/a		1		1 2	1	n/a	n/a			90 ± 1.0	99 ± 0.1 99 ± 0.1					No effect	1	
	c.631+29A>C	p.?	n/a		3								n/a	99 ± 0.1 99 ± 0.1					No effect	1	
	c.631+43G>T	p.?	n/a	n/a	2		1 2 3	1 2	n/a	n/a			n/a	100 ± 0.1 100 ± 0.0					No effect	1	
	c.631+183T>A	p.?	1		1		1						87 ± 1.6						No effect	1	

Table S4. Description of *BRCA2* exon 7 variants selected in this study.

Variations *BRCA2* Exon 7 (n = 17)	Effect on splicing	Exon 7 inclusion (%)	ΔMES (-15%)	ΔSSFL (-5%)	ΔMES and ΔSSFL	SPiCE (11,5%)
WT	-	85	-	-		-
c.517-22_-23del	↑ Skipping	60	n/a	n/a	n/a	n/a
c.517-20A>G	↑ Skipping	24	-10	n/a	n/a	n/a
c.517-19C>T	↑ Skipping	69	-2	n/a	n/a	n/a
c.517-13_-9delTCTT	↑ Skipping	0	-32	-3	1/2	42
c.517-2A>G	↑ Skipping	0	-100	-100	2/2	100
c.517-1G>A	↑ Skipping Δ7p(1nt)	0	-100	-100	2/2	100
c.517G>C	↑ Skipping	4	-12	-6	1/2	38
c.631G>T	↑ Skipping	5	-95	-100	2/2	100
c.631+2T>C	↑ Skipping	4	-100	-100	2/2	100
c.631+3A>G	↑ Skipping	5	-67	-6	2/2	96
c.631+7A>G	↑ Skipping	68	n/a	n/a	n/a	n/a
c.517-7C>T	↑ Inclusion	98	-1	2	0/2	2
c.517-4C>G	↑ Inclusion	98	11	0	0/2	1
c.631+25C>T	No effect	91	n/a	n/a	n/a	n/a
c.631+29A>C	No effect	nd	n/a	n/a	n/a	n/a
c.631+43G>T	No effect	nd	n/a	n/a	n/a	n/a
c.631+183T>A	No effect	87	n/a	n/a	n/a	n/a
True calls		Positives	6	6	5	7
		Negatives	2	2	2	2
		Total	**8**	**8**	**7**	**9**
False calls		Positives	0	0	0	0
		Negatives	3	1	2	0
		Total	**3**	**1**	**2**	**0**
Sensitivity			67	86	71	100
Specificity			100	100	100	100
Accuracy			73	89	78	100

Table S5. Comparison of variant-associated splicing effects obtained with pCAS2-*BRCA2*e7 minigenes carrying variants mapping to *BRCA2* exon 7 splice sites or to flanking intronic positions and associated splice site-dedicated *in silico* predictions.

Variations *BRCA2* Exon 7 (n = 63)	Effect on splicing	Exon 7 inclusion (%)	QUEPASA (-0.65)	HEXplorer (-28)	SPANR (+0.35%)	HAL (-5%)	LR_{skip} (33.6%)	QUEPASA & HAL	At least 3
WT	-	85	0	0	0	0	0	n/a	n/a
c.523C>T	↑ Skipping	4	-3.43	-63.7	-81.69	-36.5	98.5	2/2	4/4
c.526A>T	↑ Skipping	69	-0.62	-25.1	-8.59	-29.8	50.2	1/2	2/4
c.535C>T	↑ Skipping	73	-1.57	-68.9	-0.71	-11	56.6	2/2	4/4
c.538_539insAT	↑ Skipping	69	-1.2	-44.6	n/a	-16.7	48.7	2/2	3/4
c.544G>C	↑ Skipping	72	-0.7	-2.0	-3.09	-6.9	30.4	2/2	3/4
c.550C>G	↑ Skipping	42	-0.7	-0.7	1.59	-17.3	31.9	2/2	2/4
c.554G>T	↑ Skipping	31	-3.16	-62.3	-3.25	-72.1	91.8	2/2	4/4
c.559G>T	↑ Skipping	4	-3.39	-69.1	-82.22	-69.8	99.3	2/2	4/4
c.565G>C	↑ Skipping	65	-1.9	-29.6	-2.88	-10	52.8	2/2	4/4
c.565G>T	↑ Skipping	51	-1.87	-66.5	-0.6	-27	67.6	2/2	4/4
c.567T>G	↑ Skipping	75	-0.6	-20.0	0.49	-0.3	27.6	0/2	0/4
c.568_575del	↑ Skipping	47	-0.36	-12.5	n/a	0.6	23.9	0/2	0/4
c.569C>T	↑ Skipping	32	-1.58	-56.0	-0.64	-26	61.5	2/2	4/4
c.569C>A	↑ Skipping	41	-0.68	-12.7	1.87	-31.4	41.0	2/2	2/4
c.569_570insAT	↑ Skipping	67	-0.81	-19.7	n/a	-25.1	42.4	2/2	2/4
c.574dup	↑ Skipping	76	-1.05	-11.9	n/a	-11	36.4	2/2	2/4
c.578C>T	↑ Skipping	55	-1.41	-59.1	1.25	-16.7	53.8	2/2	3/4
c.594_598dup	↑ Skipping	55	-1.5	-9.5	n/a	-13.7	42.8	2/2	2/4
c.596C>T	↑ Skipping	62	-0.26	-55.6	-2.11	-67.6	67.3	1/2	3/4
c.598_599insGCTAC	↑ Skipping	15	-3.55	-16.3	n/a	-44	81.6	2/2	2/4
c.612_613insCTTA	↑ Skipping	58	1.9	-27.3	n/a	1.9	9.8	0/2	0/4
c.620C>G	↑ Skipping	5	-2.55	-55.3	-4.79	-63.9	87.1	2/2	4/4
c.627C>G	↑ Skipping	72	0.65	26.0	4.04	8.7	8.8	0/2	0/4
c.522T>C	No effect	83	-0.12	6.4	-1.25	5.8	17.7	0/2	1/4
c.538A>T	No effect	82	0.53	-66.0	-3.36	-6.1	29.4	1/2	3/4
c.566_567insG	No effect	80	0.24	-38.8	n/a	-9.7	26.8	1/2	2/4
c.568C>G	No effect	80	-0.23	13.0	1.20	4.9	16.9	0/2	0/4
c.585A>G	No effect	89	0.07	9.0	4.9	5.4	13.7	0/2	0/4
c.588T>C	No effect	90	-0.28	53.5	1.98	9	11.3	0/2	0/4
c.589T>G	No effect	86	-1.49	29.3	0.92	8.4	23.9	1/2	1/4
c.590C>G	No effect	81	-0.26	-13.8	2.45	-3.5	23.4	1/2	1/4
c.604C>T	No effect	83	1.09	54.4	-0.34	6.4	6.6	0/2	1/4
c.605C>T	No effect	85	1.66	12.2	-2.89	2.4	8.4	0/2	1/4
c.606C>T	No effect	82	0.17	43.8	1.35	8.4	10.2	0/2	0/4
c.609C>T	No effect	83	0.59	-1.7	-2.23	2.9	15.0	0/2	1/4
c.615_616insGAGT	No effect	15	0.03	3.1	n/a	-11.3	25.5	1/2	1/4
c.615T>G	No effect	87	-1.02	7.3	-2.76	3.2	27.4	1/2	2/4
c.622G>C	No effect	82	1.5	-13.9	-1.04	7.1	9.8	0/2	2/4
c.524A>G	↑ Inclusion	99	2.76	3	3.98	7	4.0	0/2	0/4
c.536A>G	↑ Inclusion	94	1.02	44.6	2.49	8.1	6.6	0/2	0/4
c.538A>G	↑ Inclusion	95	1.61	31.5	-2.23	6.7	6.6	0/2	1/4
c.541del	↑ Inclusion	93	1.13	53.0	n/a	4.9	6.7	0/2	0/4
c.543T>G	↑ Inclusion	97	1.24	65.4	4.67	6.3	4.8	0/2	0/4
c.545A>G	↑ Inclusion	99	1.04	46.8	2.34	9	6.3	0/2	0/4
c.548G>A	↑ Inclusion	97	0.09	-0.6	1.67	-2.7	18.5	0/2	0/4
c.548G>C	↑ Inclusion	97	0.5	9.8	0.88	-6.8	15.7	1/2	1/4
c.552A>G	↑ Inclusion	99	2.09	-9.6	0.43	-9.4	9.5	1/2	1/4
c.560A>T	↑ Inclusion	94	0.28	12.4	1.8	3.5	13.7	0/2	0/4
c.561G>T	↑ Inclusion	98	1.44	52.1	3.62	9.1	4.8	0/2	0/4
c.569_570insAACG	↑ Inclusion	98	0.31	8.3	n/a	4.9	14.2	0/2	0/4
c.573T>A	↑ Inclusion	93	0.19	10.5	-2.34	7.10	15.1	0/2	1/4
c.583del	↑ Inclusion	2	0.90	5.4	n/a	-8.2	14.3	1/2	/4
c.584C>G	↑ Inclusion	92	1.24	17.5	-80.28	8.2	50.4	0/2	1/4
c.593T>C	↑ Inclusion	100	4.09	127.8	5.18	11	0.6	0/2	0/4
c.593_596delinsAGG	↑ Inclusion	6	-0.26	-7.4	n/a	-8.4	25.5	1/2	1/4
c.595G>A	↑ Inclusion	97	1.16	-10.0	-1.51	7.6	11.1	0/2	1/4
c.595C>G	↑ Inclusion	98	2.61	28.3	1.51	8.6	3.6	0/2	0/4
c.595_598del	↑ Inclusion	98	2.38	-1.5	n/a	-0.7	6.5	0/2	0/4
c.595_599dup	↑ Inclusion	96	1.1	38.8	n/a	6.6	7.4	0/2	0/4
c.598A>G	↑ Inclusion	97	1.4	10.0	-1.79	8.4	8.3	0/2	1/4
c.605C>G	↑ Inclusion	96	2.4	37.8	2.35	6.7	3.7	0/2	0/4
c.608C>A	↑ Inclusion	92	1.78	53.0	1.6	9.9	4.2	0/2	0/4

c.610_613dup	↑ Inclusion	97	-0.68	-37.5	n/a	-21.7	43,0	2/2	3/4
True calls		Positives	17	11	11	19	17	17	12
		Negatives	37	37	19	32	38	39	38
		Total	**54**	**48**	**30**	**51**	**55**	**56**	**50**
False calls		Positives	3	3	12	8	2	1	2
		Negatives	6	12	5	4	6	6	11
		Total	**9**	**15**	**17**	**12**	**8**	**7**	**13**
Sensitivity			74	47	69	83	74	74	52
Specificity			93	93	61	80	95	98	95
Accuracy			86	76	64	81	87	89	79

Table S6. Comparison of variant-associated splicing effects obtained in the pCAS2-*BRCA2*e7 minigene assay and associated SRE-dedicated *in silico* predictions.

Variations *BRCA2* Exon 7 (n = 97)	Effect on splicing	Exon 7 skipping (%)	QUEPASA (-0.33)	HEXplorer (-11)	SPANR (+0.38%)	HAL (-4.2%)	LR_{skip} (27.5%)	QUEPASA & HAL	At least 3
WT	-	85	0	0	0	0	0	n/a	n/a
c.520C>T	↑ Skipping	10	-2.81	-54.2	-3.78	-5.5	67.8	2/2	4/4
c.521G>A	↑ Skipping	66	-1.36	8	-6.15	-3.4	36.3	1/2	2/4
c.523C>T	↑ Skipping	4	-3.43	-63.7	-81.69	-36.5	98.5	2/2	4/4
c.526A>T	↑ Skipping	69	-0.62	-25.1	-8.59	-29.8	50.2	2/2	4/4
c.535C>T	↑ Skipping	73	-1.57	-68.9	-0.71	-11	56.6	2/2	4/4
c.538_539insAT	↑ Skipping	69	-1.2	-44.6	n/a	-16.7	48.7	2/2	3/4
c.544G>C	↑ Skipping	72	-0.7	-2.0	-3.09	-6.9	30.4	2/2	3/4
c.550C>G	↑ Skipping	42	-0.7	-0.7	1.59	-17.3	31.9	2/2	2/4
c.554G>T	↑ Skipping	31	-3.16	-62.3	-3.25	-72.1	91.8	2/2	4/4
c.559G>T	↑ Skipping	4	-3.39	-69.1	-82.22	-69.8	99.3	2/2	4/4
c.559G>A	↑ Skipping	37	-2.42	-30.2	-3.36	-45.2	76.2	2/2	4/4
c.565G>C	↑ Skipping	65	-1.9	-29.6	-2.88	-10	52.8	2/2	4/4
c.565G>T	↑ Skipping	51	-1.87	-66.5	-0.6	-27	67.6	2/2	4/4
c.567T>G	↑ Skipping	75	-0.6	-20.0	0.49	-0.3	27.6	2/2	2/4
c.568_575del	↑ Skipping	47	-0.36	-12.5	n/a	0.6	23.9	1/2	2/4
c.569C>T	↑ Skipping	32	-1.58	-56.0	-0.64	-26	61.5	2/2	4/4
c.569C>A	↑ Skipping	41	-0.68	-12.7	1.87	-31.4	41.0	2/2	3/4
c.569_570insAT	↑ Skipping	67	-0.81	-19.7	n/a	-25.1	42.4	2/2	3/4
c.572A>T	↑ Skipping	72	-0.68	-49.6	-1.81	-39.1	56.5	2/2	4/4
c.574dup	↑ Skipping	76	-1.05	-11.9	n/a	-11	36.4	2/2	3/4
c.578C>T	↑ Skipping	55	-1.41	-59.1	1.25	-16.7	53.8	2/2	3/4
c.581G>A	↑ Skipping	13	-3.01	-28.7	-82.1	-54	98.3	2/2	4/4
c.587G>T	↑ Skipping	48	-0.66	-79.0	-0.77	-29.7	57.2	2/2	4/4
c.587G>A	↑ Skipping	50	-1.0	-41.7	0.9	-47.3	60.4	2/2	3/4
c.594_598dup	↑ Skipping	55	-1.5	-9.5	n/a	-13.7	42.8	2/2	2/4
c.596C>T	↑ Skipping	62	-0.26	-55.6	-2.11	-67.6	67.3	1/2	3/4
c.599C>A	↑ Skipping	46	-1.07	-51.3	-0.89	-37.8	60.4	2/2	4/4
c.599C>T	↑ Skipping	41	-0.97	-54.9	2	-65.6	71.3	2/2	3/4
c.598_599insGCTAC	↑ Skipping	15	-3.55	-16.3	n/a	-44	81.6	2/2	4/4
c.612_613insCTTA	↑ Skipping	58	1.9	-27.3	n/a	1.9	9.8	0/2	1/4
c.617C>G	↑ Skipping	25	-1.12	-32.8	0.68	-64.1	68.8	2/2	4/4
c.620C>G	↑ Skipping	5	-2.55	-55.3	-4.79	-63.9	87.1	2/2	4/4
c.620C>T	↑ Skipping	21	-1.94	-54.7	0.32	-75.8	84.3	2/2	4/4
c.623T>G	↑ Skipping	75	0.16	-38.6	0.22	-4.5	25.2	1/2	3/4
c.627C>G	↑ Skipping	72	0.65	26.0	4.04	8.7	8.8	0/2	0/4
c.522T>C	No effect	83	-0.12	6.4	-1.25	5.8	17.7	0/2	1/4
c.522T>G	No effect	86	-0.31	-2.4	1.32	4.2	20.2	1/2	0/4
c.534A>G	No effect	83	0.75	27.9	0.47	6.9	9.4	0/2	0/4
c.538A>T	No effect	82	0.53	-66.0	-3.36	-6.1	29.4	0/2	2/4
c.556G>C	No effect	84	1.44	-21.5	2.28	-6.2	12.6	1/2	1/4
c.566_567insG	No effect	80	0.24	-38.8	n/a	-9.7	26.8	1/2	2/4
c.568C>G	No effect	80	-0.23	13.0	1.20	4.9	16.9	0/2	0/4
c.572delinsCT	No effect	84	-0.63	-27.4	n/a	-3.9	31.3	1/2	3/4
c.573T>C	No effect	88	0.46	27.7	3.07	6.4	10.2	0/2	0/4
c.582G>A	No effect	84	0.33	25.4	-80.64	3.4	62.9	0/2	1/4
c.585A>G	No effect	89	0.07	9.0	4.9	5.4	13.7	0/2	0/4
c.588T>C	No effect	90	-0.28	53.5	1.98	9	11.3	1/2	1/4
c.589T>G	No effect	86	-1.49	29.3	0.92	8.4	23.9	1/2	1/4
c.590C>G	No effect	81	-0.26	-13.8	2.45	-3.5	23.4	0/2	1/4
c.601C>G	No effect	87	0.44	12.6	0.82	5.7	12.5	0/2	0/4
c.602C>G	No effect	84	2.05	11.3	8.19	-9.2	6.4	1/2	1/4
c.604C>T	No effect	83	1.09	54.4	-0.34	6.4	6.6	0/2	1/4
c.605C>T	No effect	85	1.66	12.2	-2.89	2.4	8.4	0/2	1/4
c.606C>T	No effect	82	0.17	43.8	1.35	8.4	10.2	0/2	0/4
c.609C>T	No effect	83	0.59	-1.7	-2.23	2.9	15.0	0/2	1/4
c.610delC	No effect	82	0.64	23.8	n/a	9.4	9.9	0/2	0/4
c.615T>G	No effect	87	-1.02	7.3	-2.76	3.2	27.4	1/2	2/4
c.615_616insGAGT	No effect	15	0.03	3.1	n/a	-11.3	25.5	1/2	1/4
c.622G>C	No effect	82	1.5	-14.2	-1.04	7.1	9.8	0/2	2/4
c.625C>T	No effect	82	1.16	-49.9	1.34	2.7	15.6	0/2	1/4
c.627C>A	No effect	81	0.65	26.0	4.04	8.7	8.8	0/2	0/4
c.627C>T	No effect	87	-1.37	-33.7	2.40	3.4	36.0	1/2	2/4
c.524A>G	↑ Inclusion	99	2.76	3	3.98	7	4,0	0/2	0/4

c.532A>C	↑ Inclusion	97	1.05	20.4	-3.47	7.4	9.6	0/2	1/4
c.536A>G	↑ Inclusion	94	1.02	44.6	2.49	8.1	6.6	0/2	0/4
c.538A>G	↑ Inclusion	95	1.61	31.5	-2.23	6.7	6.6	0/2	1/4
c.538_539del	↑ Inclusion	93	1.61	35.5	n/a	7	5.9	0/2	0/4
c.539T>C	↑ Inclusion	97	1.93	70.7	-1.38	10	3.6	0/2	1/4
c.541del	↑ Inclusion	93	1.13	53.00	n/a	4.9	6.7	0/2	0/4
c.543T>G	↑ Inclusion	97	1.24	65.4	4.67	6.3	4.8	0/2	0/4
c.545A>G	↑ Inclusion	99	1.04	46.8	2.34	9	6.3	0/2	0/4
c.548G>A	↑ Inclusion	97	0.09	-0.6	1.67	-2.7	18.5	0/2	0/4
c.548G>C	↑ Inclusion	97	0.5	9.8	0.88	-6.8	15.7	1/2	1/4
c.549T>C	↑ Inclusion	94	-0.23	15.1	0.90	6.0	16.4	0/2	0/4
c.551T>C	↑ Inclusion	99	1.91	86.6	2.54	10.5	2.8	0/2	0/4
c.552A>G	↑ Inclusion	99	2.09	-9.6	0.43	-9.4	9.5	1/2	1/4
c.560A>T	↑ Inclusion	94	0.28	12.4	1.8	3.5	13.7	0/2	0/4
c.561G>T	↑ Inclusion	98	1.44	52.1	3.62	9.1	4.8	0/2	0/4
c.569_570insAACG	↑ Inclusion	98	0.31	8.3	n/a	4.9	14.2	0/2	0/4
c.573T>A	↑ Inclusion	93	0.19	10.5	-2.34	7.10	15.1	0/2	1/4
c.574_575del	↑ Inclusion	96	0.94	43.9	n/a	8.9	7.3	0/2	0/4
c.575T>C	↑ Inclusion	100	1.4	54.9	6.05	10.8	4.3	0/2	0/4
c.583del	↑ Inclusion	2	0.90	5.4	n/a	-8.2	14.3	1/2	1/4
c.584C>G	↑ Inclusion	92	1.24	17.5	-80.28	8.2	50.4	0/2	1/4
c.593T>C	↑ Inclusion	100	4.09	127.8	5.18	11	0.6	0/2	0/4
c.593_596delinsAGG	↑ Inclusion	6	-0.26	-7.4	n/a	-8.4	25.5	1/2	1/4
c.595G>A	↑ Inclusion	97	1.16	-10.0	-1.51	7.6	11.1	0/2	1/4
c.595C>G	↑ Inclusion	98	2.61	28.3	1.51	8.6	3.6	0/2	0/4
c.595_598del	↑ Inclusion	98	2.38	-1.5	n/a	-0.7	6.5	0/2	0/4
c.595_599dup	↑ Inclusion	96	1.1	38.8	n/a	6.6	7.4	0/2	0/4
c.598A>G	↑ Inclusion	97	1.4	10.0	-1.79	8.4	8.3	0/2	1/4
c.605C>G	↑ Inclusion	96	2.4	37.8	2.35	6.7	3.7	0/2	0/4
c.608C>A	↑ Inclusion	92	1.78	53.0	1.6	9.9	4.2	0/2	0/4
c.610_613dup	↑ Inclusion	97	-0.68	-37.5	n/a	-21.7	43.0	2/2	3/4
c.619A>G	↑ Inclusion	93	0.40	6.2	1.65	7.1	12.8	0/2	0/4
True calls	Positives		31	30	21	30	31	29	28
	Negatives		56	53	30	54	56	56	60
	Total		**87**	**83**	**51**	**84**	**87**	**85**	**88**
False calls	Positives		6	9	15	8	6	1	2
	Negatives		4	5	9	5	4	6	7
	Total		**10**	**14**	**24**	**13**	**10**	**7**	**9**
Sensitivity			89	86	70	86	88	81	80
Specificity			90	85	67	87	90	98	97
Accuracy			90	86	68	87	90	92	91

Table S7. Comparison of variant-associated splicing effects obtained in the pCAS2-*BRCA2*e7 minigene assay and associated SRE-dedicated *in silico* predictions after optimization of the thresholds.

MLH1 exon 7, an emblematic exon sensitive to intronic mutations but not to alterations of exonic splicing regulators, sheds light into the performance of SRE-dedicated bioinformatics approaches

Rolain M[1*], Tubeuf H[1,2*], Lanzing F[1], Soukarieh O[1], Polo P[1], Drouet A[1], Frebourg T[3], Martins A[1]

1. Inserm-U1245, UNIROUEN, Normandie University, Normandy Centre for Genomic and Personalized Medicine, Rouen, France, **2** Interactive Biosoftware, Rouen, France, **3** Department of Genetics, University Hospital, Normandy Centre for Genomic and Personalized Medicine, Rouen, France

* These authors contributed equally to this work

Author contributions

AM conceived and designed the project. MR, FL and HT generated the experimental data. HT performed the bioinformatics and statistical analyses. HT, MR and AM were involved in data interpretation. HT and AM wrote the manuscript with input from MR. All authors read and approved the final manuscript.

Acknowledgments

This project was supported by the OpenHealth Institute, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (Gefluc) as well as the European Union and Région Normandie. Europe gets involved in Normandie with European Regional Development Fund (ERDF). HT was funded by a CIFRE PhD fellowship (#2015/0335) from the French Association Nationale de la Recherche et de la Technologie in the context of public-private partnership between INSERM and Interactive Biosoftware.

Abstract

Alterations of cis-elements implicated in RNA splicing play a major role in the development of many genetic disorders but are often difficult to predict. Recently, a class of 86 disease-causing genes was identified as being particularly enriched in mutations at 3' and 5' splice sites (3'/5'ss), including *MLH1* a gene implicated in hereditary cancer. Moreover, certain MLH1 exons (8, 10, and 15) were shown to be very sensitive to genetic variations that alter potential exonic splicing regulatory elements (SREs) suggesting that either exonic SRE mutations are underestimated in many exons/genes or that these *MLH1* exons are exceptionally vulnerable to SRE alterations. Here, we investigated the impact on splicing of 63 variants reported in the exon 7 of *MLH1* (n=32) and flanking intronic sequences (n=31). Predictions of splicing defects were performed with computational methods dedicated to 3'/5'ss, branchpoints and SREs, whereas actual RNA splicing outcomes were assessed by cell-based minigene reporter assays. Our results revealed a total of 15 splicing mutations: 13 within the 3'/5'ss, one in intron 7 likely affecting an SRE, and one that created a new 5'ss within the exon. Importantly, we did not detect splicing defects resulting from alterations in exonic SREs. Contrary to splice-site related predictions, those dedicated to exonic SREs produced an unexpected high number of false calls. Additional analyses revealed that the resistance of exon 7 to SRE mutations was due to the important strength of its 3'/5'ss. Importantly, replacing these sites by suboptimal counterparts restored the predictive power of SRE-dedicated tools. In sum, we demonstrate that not all exons are sensitive to SRE mutations and that the strength of 3'/5'ss can influence the performance of SRE-dedicated prediction tools. These findings may be useful for improving bioinformatics strategies aiming at predicting which variants are more likely to alter RNA splicing by affecting exonic SREs.

Keywords

RNA splicing, Lynch syndrome, bioinformatics predictions, splicing regulatory elements, branchpoint, molecular diagnostics

Introduction

Recent advances in high-throughput methods made sequencing of human genomes and exomes largely widespread leading to an increasing repertoire of human genetic variation (Rabbani *et al.*, 2014; Shendure, 2011). As a result, millions of nucleotide changes have been identified in patients as well as in healthy individuals, many detected in genes already implicated in disease (Lek *et al.*, 2016). Despite the creation of databases facilitating access to variant-related data, such as ClinVar, LOVD and gnomAD, a large fraction of variations remain of uncertain pathogenicity (also called VUS for variants of unknown significance) or of conflicting interpretation (Eilbeck *et al.*, 2017; Landrum *et al.*, 2014; Plazzer *et al.*, 2013). VUS are not actionable and thus represent a major challenge in medical genetics and an obstacle to the progress of precision medicine (Cooper and Shendure, 2011; Frebourg, 2014). The clinical classification of human variants is a complex task that depends on multiple pieces of evidence including, among others, functional data relative to potential biological consequences at the molecular or cellular levels (Richards *et al.*, 2015). For practical reasons, this information is lacking for most variants today. Nevertheless, there is growing awareness that intragenic variants, either intronic or exonic, can cause or modify a disease phenotype by altering cis-acting elements important for proper pre-mRNA splicing (Baralle and Buratti, 2017). These elements include core signals that directly participate in the splicing reactions, notably the 3' and 5' splice sites (3'ss and 5'ss) and the branch points (BPs), as well as auxiliary signals known as splicing regulatory elements (SREs) that include both exonic and intronic splicing regulators (ESR and ISR, respectively) (Wang and Burge, 2008). Depending on whether they promote or inhibit exon inclusion, SREs can be classified as enhancers (ESE and ISE) or silencers (ESS and ISS). Even if theoretically any intragenic VUS is susceptible of altering one of these signals, it is currently impractical to experimentally testing all variants to verify their potential impact on splicing. Computational predictive methods became thus of paramount importance for establishing priorities for functional testing. This is especially the case for *in silico* tools that focus on 3'ss and 5'ss, such as MaxEntScan, SpliceSiteFinder-Like and HSF, which have shown compelling predictive power in several studies and are now routinely used by many molecular diagnostic laboratories (Houdayer *et al.*, 2012; Leman *et al.*, 2018; Moles-Fernández *et al.*, 2018; Soukarieh *et al.*, 2016). In contrast, those focusing on BPs and on SREs are seldom used, probably because these splicing signals are more difficult to define and the predictive performances

of the corresponding dedicated *in silico* tools have not been be widely evaluated (Baralle and Buratti, 2017; Grodecká *et al.*, 2017; Mercer *et al.*, 2015; Soukarieh *et al.*, 2016).

Today, at least 15%-35% of all disease-causing mutations are known to cause aberrant RNA splicing (Baralle *et al.*, 2009; Lim *et al.*, 2011; Mort *et al.*, 2014), with some studies reaching estimates of up to 50% splicing mutations in certain genes, such as *ATM* (Teraoka *et al.*, 1999) and *NF1* (Ars *et al.*, 2000). Whether some genes are particularly prone to splicing mutations or if this type of alterations is still under recognized in many disease-causing genes, remains an important question (López-Bigas *et al.*, 2005). Recently, a study based on a comprehensive analysis of all disease-causing splicing mutations reported in the Human Genome Mutation Database (HGMD) pointed to the existence of a group of 86 genes especially affected by splice-site mutations (SSM) in which cancer genes were overrepresented. These included not only *ATM* and *NF1* but also the mismatch repair (*MMR*) gene MLH1 among others (Rhine *et al.*, 2018). *MLH1* is implicated in Lynch syndrome (LS), one of the most frequent forms of autosomal inherited predispositions to cancer, notably colorectal and endometrium cancers (Thompson *et al.*, 2014). Importantly, *MLH1* exhibits a large mutational spectrum, with at least 30% of variants currently classified as VUS (Grandval *et al.*, 2013; Thompson *et al.*, 2014). Over the last few years, several experimental studies based either on patients' RNA analysis and/or minigene reporter assays revealed that numerous *MLH1* VUS alter RNA splicing (Auclair *et al.*, 2006; van der Klift *et al.*, 2015; Lastella *et al.*, 2006; Tournier *et al.*, 2008). In particular, our functional analysis of all variants reported in *MLH1* exon 10 unveiled a especially vulnerable exon, with a fraction of splicing mutations higher than expected (77%, 17/22), most of them mapping to positions outside the splice sites and thus likely affecting ESRs (Soukarieh *et al.*, 2016). Moreover, we discovered that the impact on splicing of most *MLH1* exon 10 variants could be accurately predicted by *in silico* tools, notably by using "classic" splice site-dedicated methods and newly developed SRE-dedicated bioinformatics approaches (Soukarieh *et al.*, 2016). The promising character of the SRE-dedicated approaches (QUEPASA, HEXplorer, SPANR and HAL) was further confirmed by a recent retrospective large scale study that included more than 1200 variants within 87 different genes followed by a prospective analysis of 150 variants mapping to 3 different exons (Tubeuf *et al.*, in preparation). The performances of SRE-dedicated methods seemed however to depend on the gene/exon under consideration, suggesting that additional analyses should be performed to better understand the applicability and limitations of these approaches. Moreover, understanding not only which genes

but also which exons are more sensitive to splicing mutations would also be useful for helping stratifying variants for functional analysis. For these reasons, we now decided to investigate the exon 7 of *MLH1* for which 4 single nucleotide variants (SNVs) were recently tested in a minigene assay, none showing an impact on splicing (Rhine *et al.*, 2018). These results suggested that exon 7 was less affected by splicing mutations than other MLH1 exons, such as exons 8, 10 and 15 (Rhine *et al.*, 2018), but the number of analyzed variants was too small to reach a definitive conclusion.

Here we report a comprehensive computational and experimental analysis of a large number of variants mapping to *MLH1* exon 7 and adjacent intronic positions. Our goal was to evaluate the incidence of splicing mutations in this genetic region and concomitantly pursue an assessment of the performance of splicing-dedicated *in silico* tools, especially of QUEPASA, HEXplorer, SPANR and HAL. We found a high number of splicing mutations at intronic positions but only one within the exon. Additional analyses showed that the resistance of *MLH1* exon 7 to potential ESR mutations was due to the strengths of the 3'ss and 5'ss, and that this feature interfered with the predictive power of the SRE-dedicated approaches.

Material & Methods

Nomenclature. Nucleotide numbering is based on the cDNA sequence of *MLH1*, RefSeq NM_000249.3, with c.1 denoting the first nucleotide of the translation initiation codon, as recommended by the Human Genome Variation Society.

Retrieval of *MLH1* exon 7 variants. We collected all nucleotide variants reported in *MLH1* exon 7 and in its proximal flanking intronic sequences (c.546-40 to c.588+40), until June 2017, by interrogating 8 different human variation databases, all of which, except for UMD and LOVD, were accessed via the integrated software Alamut Visual (Version 2.11, Interactive Biosoftware, http://www.interactive-biosoftware.com). More specifically, variants were compiled from the following databases: UMD-MLH1 (Grandval *et al.*, 2013) (Universal Mutation Database-*MLH1*, http://www.umd.be/MLH1/), LOVD (Plazzer *et al.*, 2013) (Leiden Open Variation Database, http://chromium. lovd.nl/LOVD2/colon_cancer/home.php?select_ db=MLH1), ClinVar (Lek *et al.*, 2016) (https://www.ncbi.nlm. nih.gov/clinvar/), HGMD (Human Gene Mutation Database,

http://www.hgmd.cf.ac.uk/ac/index. php), COSMIC (Catalogue of Somatic Mutations in Cancer, http://cancer.sanger.ac.uk/cosmic), dbSNP (the Single Nucleotide Polymorphism database http://www.ncbi.nlm.nih.gov/SNP/), gnomAD (Exome Aggregation Consortium, http://exac. broadinstitute.org/) and ESP (Exome Sequencing Project, http://evs.gs.washington.edu/EVS/).

Splicing minigene reporter assays. In order to evaluate the impact on RNA splicing of the selected *MLH1* exon 7 variants, we performed a functional assay based on the comparative analysis of the splicing pattern of wild-type (WT) and mutant reporter minigenes, as follows. Minigenes were prepared by using the pCAS2 vector (Soukarieh *et al.*, 2016), ensuing previously described procedures (Gaildrat *et al.*, 2010) with a few modifications. The WT genomic fragments *MLH1* c.546-157_c.677+231 (*MLH1* exons 7-8 and flanking intronic sequences) were inserted into the BamHI and MluI cloning sites of the reporter plasmid pCAS2, yielding the four-exon hybrid minigene pCAS2-*MLH1e7*-8 (Figure S1). Nucleotide variants were introduced by site-directed mutagenesis by using the two-stage overlap extension PCR method (Ho *et al.*, 1989), and primers indicated in Table S1, except c.588+5G>A and c.588+11G>C that were sub-cloned from already available pCAS1 constructs (Tournier *et al.*, 2008), and c.588+5G>T that was amplified from patient gDNA as recently described (Piňero *et al.*, in preparation). Then, the mutant amplicons were introduced by homologous recombination using the SLICE method (Motohashi, 2015) into the pCAS2 vector previously digested with BamHI and MluI. All constructs were sequenced to ensure that no unwanted mutations had been introduced into the inserted fragments during PCR or cloning. Next, WT and mutant minigenes (400 ng/well) were transfected in parallel into HeLa cells grown at ~70% confluence in 12-well plates using the FuGENE 6 transfection reagent (Roche Applied Science). HeLa cells obtained from ATCC were cultivated in Dulbecco's modified Eagle medium (Life Technologies) supplemented with 10% fetal calf serum in a 5% CO2 atmosphere at 37°C. Twenty-four hours later, total RNA was extracted using the NucleoSpin RNA II kit (Macherey Nagel) according to the manufacturer's instructions, and the minigenes' transcripts were analysed by semi-quantitative fluorescent RT-PCR (25 cycles of amplification) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng total RNA and the 6FAM-pCASKO1F and pCAS-2R minigene primers (Table S1). RT-PCR products were separated by electrophoresis on 2.5% agarose gels containing ethidium bromide and visualized by exposure to ultraviolet light under saturating conditions using the Gel Doc XR image acquisition system (Bio-RAD), followed by gel-purification and sanger sequencing for proper identification of the

minigene's transcripts. In parallel, splicing events were quantitated by performing capillary electrophoresis on an automated sequencer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) followed by a computational analysis with the GeneMapper v5.0 software (Applied Biosystems).

Branch point mapping. First, total RNA was extracted from a human lymphoblastoid cell line obtained from a healthy individual. Then, we incubated 1 µg DNase-treated RNA with RNase R (30U, Epicentre) for 30 min at 37°C. After purification, cDNA synthesis was performed with 500 ng RNase R-treated RNA using SuperScriptIITM Reverse Transcriptase (Life Technologies) and an outer reverse primer (Table S1). The first round of PCR (35 cycles of amplification) was set up using cDNA and FIREPol® DNA polymerase (SOLIS BIODYNE) with an outer primer set (Table S1), and divided into multiple reactions performed at different annealing temperatures (50-60°C). PCR products were pooled and purified. Purified DNA was used as template for a second round of PCR (35 cycles of amplification) using an inner primer set, divided across different annealing temperatures (50-60°C). PCR products were combined and run on an agarose gel. Bands of interest were excised, DNA extracted and purified and the products were ligated into the pGemTEasy cloning vector (Promega). Plasmid DNA from 10 colonies were Sanger sequenced using a T7 Promoter primer.

Splicing-dedicated bioinformatics predictions. Three types of bioinformatics methods were used to predict variant-induced splicing alterations, namely: splice site (ss)- , branch point (BP)- or splicing regulatory elements (SRE)- dedicated methods depending on the position of the variants relative to the exon. For intronic variants and for those mapping at exon termini on positions overlapping the splice sites, we resorted to MaxEntScan (Yeo and Burge, 2004) (MES, http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html; MaximumEntropy Model), SpliceSiteFinder-like (Shapiro and Senapathy, 1987) (SSFL, http://www.interactive-biosoftware.com) and Splicing Prediction in Consensus Elements (SPiCE) (Leman *et al.*, 2018). MES and SSFL were used both as stand-alone and in a combined sequential fashion (MES+SSFL) as recommended by Houdayer and colleagues (Houdayer *et al.*, 2012). These algorithms were interrogated by using either the integrated software tool Alamut Batch v1.9 or Alamut Visual v2.11 (Interactive Biosoftware, http://www.interactive-biosoftware.com) whereas SPiCE scores were retrieved from the dedicated software tool (https://sourceforge.net/projects/spicev2-1/). Moreover,

to predict 3'ss and 5'ss strengths of all MLH1 exons in the WT context, we also relied on MES and SSFL scores as well as on the frequencies of exon-intron junctions provided by Alamut Visual v2.11, which automatically determines the occurrence of each 6 mer and 9 mer among 10728 human 3' and 5' splice sites (4 intronic and 2 exonic nucleotides, and 3 exonic and 6 intronic nucleotides), respectively. Alamut Visual reports the frequency of 6 mer sequences only if at least 6 out of the 8 upstream nucleotides are pyrimidines. In addition, to get a better appreciation of MLH1 splice site strengths we used Splicing Regulation Online Graphical Engine (SROOGLE, http://sroogle.tau.ac.il/), a resource that assigns percentile values to 3'ss and 5'ss MES scores relative to their ranking in a dataset of more than 50,000 constitutive or alternative exons (Schwartz *et al.*, 2009).

The position of putative branch points, as well as the potential impact of variants mapping at these sites was predicted by using Human Splicing Finder (Desmet *et al.*, 2009) (HSF, http://www.umd.be/HSF/) and SpliceSiteFinder-like (SSFL, http://www.interactive-biosoftware.com) both interrogated by using the integrated software tool Alamut Visual v 2.10 (Interactive Biosoftware, France). In addition, we also used SROOGLE (http://sroogle.tau.ac.il/, Schwartz *et al.*, 2009) that provides branch site scores based on Kol *et al.* (Kol *et al.*, 2005) and on Schwartz *et al.* (Schwartz *et al.*, 2009).

For the prediction of variant-induced impact on ESR, we resorted to four newly developed SRE-dedicated *in silico* approaches: (i) the QUEPASA method previously described by Ke and co-workers (Ke *et al.*, 2011) and implemented by our group (Di Giacomo *et al.*, 2013), which is based in the calculation of total ESRseq score changes (ΔtESRseq) (ii) the HEXplorer method (Erkelenz *et al.*, 2014) which calculates ΔHZEI values (iii) the SPANR approach described by Xiong and co-workers (Xiong *et al.*, 2015) which yields ΔΨ scores, and (iv) HAL based on the calculation of ΔΨ scores, as described by Rosenberg and co-workers (Rosenberg *et al.*, 2015)., Both ΔtESRseq and ΔHZEI scores were calculated by using the Alamut Batch prototype tool version 1.5.2 (ESRseq), (Interactive Biosoftware, http://www.interactive- biosoftware.com), whereas SPANR and HAL scores were retrieved from the dedicated online interfaces (http://tools.genes. toronto.edu and http://splicing.cs.washington. edu/ SE, respectively). For each SRE-dedicated *in silico* tool, score changes (Δ) of the exonic variants, smaller than the indicated thresholds were considered predictive

of increased exon skipping. Combinations of SRE-dedicated methods were performed as recently described (Tubeuf *et al.*, in preparation).

Statistical analyses. Results are presented as the mean ± SEM of three independent experiments. Data derived from confrontation of experimental and *in silico* analyses were compared by using either one-way ANOVA test and Pearson's correlation coefficient or their derivatives depending on data distribution patterns as detailed in Table S3. In general terms, the Kruskal-Wallis or ANOVA tests followed by Duns or Bonferroni post-tests, respectively, were used for assessing the performance of the bioinformatics tools in discriminating 3 groups of variants (i.e. variants that increase exon skipping versus those with no effect on splicing versus those that increase exon inclusion). Linear Correlation between exon inclusion levels and *in silico* predictions was measured by calculating Spearman or Pearson correlation coefficients (r). All statistical analysis were performed by using GraphPad Prism software (Version 5.0). Results are expressed as two sided p-values (* p-value<0.05, ** p-value<0.01, *** p-value<0.001) and were considered significant when p-value <0.05).

Performance assessment. The evaluation of the predictive power of splicing-dedicated bioinformatics methods was performed by measuring sensitivity (Sen) = [TPx100/(TP + FN)], specificity (Sp) = [TNx100/(TN + FP)], accuracy (Acc) = [(TN+TP)x100/(TN+TP+FN+FP), where TP (true positive) and FN (false negative) values are the numbers of positive samples that are predicted to be positive and negative respectively. Analogously, TN (true negative) and FP (false positive) values are the numbers of negative samples that are predicted to be negative and positive respectively. TP, TN, FP, FN were determined by taking into account thresholds determined either previously (Tubeuf *et al.*, in preparation), as indicated. The predictive power of the SRE-dedicated tools were then compared to each other by using Venn diagrams plotted by Jvenn (Bardou *et al.*, 2014), an interactive web application (http://jvenn.toulouse.inra.fr/app/example.html).

Results

Selection of variants located within *MLH1* exon 7 and its flanking intronic regions. We began by interrogating national and international human variation databases in order to retrieve all genetic

changes reported within *MLH1* exon 7 or in its flanking intronic regions (c.546-40 to c.588+40) either in the genome of patients suspected of Lynch syndrome, in tumors or in the general population (Table 1). As a result, we collected a total of 63 natural variations, including 55 SNVs and 8 small deletions, most of which identified in cancer patients suspected of Lynch syndrome (Table 1). Only 20 of these variations are currently classified as clearly pathogenic or likely pathogenic (class 5 or 4, respectively) and 7 as clearly not pathogenic or likely not pathogenic (class 1 or 2, respectively), whereas the remaining 36 variations include 10 VUS (class 3), 12 variations with conflicting classifications and 14 variations not yet classified (Table 1). Altogether, this selection comprises 31 intronic variations (including 1 variant overlapping an exon-intron junction) and 32 exonic changes. We then decided to determine which fraction could affect exon 7 inclusion by modifying splicing signals, either the 3' or 5' splice sites, the branch site or splicing regulatory elements. We assessed the potential impact on splicing of the variants by both performing splicing-dedicated *in silico* analyses and minigene-based assays (Table 1).

A large fraction of naturally occurring intronic variants flanking *MLH1* exon 7 affect its splicing pattern.

By using splice site-dedicated *in silico* tools (MES, SSFL, MES+SFL and SPiCE), we found that, depending on the method, 12 to 13 (39 to 42%) out of the 31 intronic variants were predicted by to induce splicing defects either by destroying or by decreasing splice site strength (Table S3). Of note, MES, SSFL, MES+SFL and SPiCE were not able to provide scores for about half of the variants, since those variants were located outside their corresponding prediction windows (Table S3).

Next, to assess the actual impact of the 31 intronic variants on *MLH1* exon 7 splicing, we performed a cell-based splicing assay with pCAS2-*MLH1*e7+8-derived minigenes. As shown on Figure 1B, the wild-type (WT) minigene generated an unique transcript containing exons 7 and 8 (100% exon inclusion), which is in agreement with the absence of alternative splicing reported for these exons (Thompson *et al.*, 2015). Importantly, the minigene assay results revealed that 14 out of the 31 intronic variations (45%) altered the splicing pattern of exon 7 relative to wild-type (Figure 1B). More precisely, 13 variants induced very drastic splicing defects (<5% exon inclusion), either exon

7 skipping (n=12) or a deletion of the last 17 nucleotides of the exon due to the usage of a de novo 5'sscreated by c.574_588+2del (Figure S2A). Only one variation induced a partial defect (~10% exon skipping, c.588+38G>A). Curiously, the variation c.546-2A>C led not only to increased exon 7 skipping (95% exon skipping) but also to deletion of the first 8 nucleotides of the exon in a small fraction of the minigene transcripts (5%). This anomaly was due to the creation of a de novo 3'ss, an effect that was not observed for c.546-2A>G nor c.546-2A>T even though a cryptic 3'ss at the same position was equally predicted for these variants (Figure S2B). Similarly, c.546-1G>A led to total exon 7 skipping despite the creation of a cryptic 3'ss being predicted as concomitant to the destruction of the natural 3'ss (Figure S2C). These cases highlight the importance of experimentally analyze predicted variant-induced splicing alterations to assess their actual outcomes. Of note, 10 of the 14 variant-induced splicing defects are currently classified as pathogenic, 7 variations corresponding now to new experimentally demonstrated splicing mutations (Table 1).

The 14 variants leading to exon 7 skipping can be separated into three categories according to their position relative to nearest reference splice site: (i) a first category consisting of 4 variants located on the 2 most conserved positions of the 3' splice site (c.546-2A>C, c.546-2A>G, c.546-2A>T and c.546-1G>A), (ii) a second category consisting of 8 variants mapping to the 5' splice site (c.588+1del, c.588+1G>T, c.588+2T>A, c.588+2T>C, c.588+3_+6del, c.588+5G>A, c.588+5G>C, c.588+5G>T and c.574_588+2del) and (iii) and, surprisingly, a third category consisting of 1 variant located in intron 7 at a distance from 5' splice site (c.588+38G>A). These results confirm the main splicing events detected on patients' RNA for 4 variations from our collection (c.546-2A>C, c.588+1G>T, c.588+5G>A, and c.588+11G>C, Table 1) and demonstrate the physiological pertinence of the pCAS2-*MLH1*e7+8 -based minigene assays.

We then assessed the performance of the splice site -dedicated *in silico* approaches in predicting *MLH1* exon 7 intronic splice site alterations. Based on the sequence window taken into account by these tools, we observed that 18/18 and 17/17 variants (100% accuracy) were correctly predicted by MES and by SPiCE, respectively; 16/17 variants (94% accuracy) were correctly predicted by the combination of MES+SSFL and 15/17 variants (88% accuracy) were correctly predicted by SSFL (Figure 1C & Table S3). Statistical analyses further highlighted the good performance of MES, SSFL, MES+SSFL and SPiCE for discriminating variants that lead to splicing alterations

(T-test: p-values <0.01). We surmise that a large fraction of intronic variants mapping near *MLH1* exon 7 have a drastic impact on splicing by directly affecting its 3'ss or 5'ss, and that this alterations can be correctly predicted by MES+SSFL and SPiCE.

MLH1 exon 7 may have redundant branch points.

In addition to its intrinsic strength, 3′ss selection also depends on additional signals including the branch point (BP) sequence. However, to date, only over a dozen disease-associated branch site mutations have been published (Lewandowska, 2013). We noticed from our minigene results that from the 13 variants tested in intron 6, only those directly mapping to the most conserved 3'ss positions affected splicing, none of the upstream variants showing an effect. To gain more insight into the dependence of *MLH1* exon 7 on intronic splicing signals, we next decided to map the BP of this exon. Of note, to our knowledge neither a BP nor a BP mutation has thus far been identified in *MLH1*.

We started by analyzing the sequence immediately upstream the 3'ss of *MLH1* exon 7 (100 nt window) by using three computational BP predictors (HSF, SSFL and SROOGLE). As shown on Figure 2A, 5 putative BPs were indicated by at least two predictors, namely c.546-23A, c.546-57A, c.546-70A, c.546-83A and c.546-93A. Then, we individually mutated each putative BP from A to C, the less frequently used BP nucleotide in constitutive introns (Pineda and Bradley, 2018) and evaluated the splicing outcomes by using the pCAS2-MLH1.ex7+8 minigene reporter assay. Despite predictions of BP destruction alterations by BP predictors (Figure 2B), the minigene assay revealed that none of the 5 mutations effected splicing (Figure 2C).

Next, we decided to experimentally identify the exact position of the BP used for *MLH1* exon 7 splicing by a mapping approach consisting in the amplification of the corresponding RNA lariat intermediate from a human cell line (Figure 2D). Sequencing of the lariat product obtained by nested RT-PCR (~ 250 bp) implicated adenine at position c.546-23 as a real BP (Figure 2E). Detecting a A-to-T mismatch at the BP upon sequencing the lariat can be considered as a sign of a correctly inferred branch point given that reverse transcriptases often incorporate an incorrect nucleotide at the 5' splice site-BP junction (Pineda and Bradley, 2018). This result is in agreement

with the proximal BP predicted by the 3 *in silico* tools. In addition, our results are consistent with current knowledge derived from annotations of human BPs, especially: (i) the position of the BP relative to the nearest 3'ss (within a window of 40 nt upstream the 3'ss for 90% of human BPs), (ii) the type of nucleotide used as BP (an adenosine for 92% of human BPs) and (iii) the consensus motif of the BS (yUnAy), including a canonical uridine at position −2 relative to the BP adenine (Gao *et al.*, 2008). Because a A-to-C transversion at the BP can sometimes be tolerated by the spliceosome (Královicová *et al.*, 2006), we then decided to bioinformatically assess the impact of all nucleotide changes at position -23 and analyze the splicing outcomes again by performing a minigene splicing assay (Figure 2E). As shown in Figures 2C and 2E, the three mutations predicted to alter the BP (c.548-23A>C, c.548-23A>G and c.548-23A>T) did not have an impact on splicing, indicating that the c.548-23A BP is not essential for efficient splicing of *MLH1* exon 7 and suggesting that other BPs may exist that compensate the loss of c.548-23A at least in the context of our minigene assay. This hypothesis is in agreement with recent data demonstrating that most human introns contain multiple branchpoints (Pineda and Bradley, 2018).

***MLH1* exon 7 is resistant to alterations of potential exonic splicing regulatory elements.**

We then asked if the splicing pattern of *MLH1* exon 7 was sensitive to naturally occurring exonic variations, a behavior that we recently observed, in an exacerbated way, for *MLH1* exon 10 (Soukarieh *et al.*, 2016). Because newly developed SRE-dedicated *in silico* approaches (QUEPASA, HEXplorer, SPANR and HAL) could correctly predict the impact on splicing of most *MLH1* exon 10 variants (Soukarieh *et al.*, 2016; Tubeuf et al, in preparation), we decided to also evaluate the performance of these tools to predict the impact on splicing of the 32 exonic variants described in *MLH1* exon 7, including 16 presumed missense, 5 nonsense, 7 synonymous and 4 frameshift variants (Table 2 and Figure 3A). As shown in Table S4, and depending on the *in silico* tool taken into account, 6 to 15 out of these 32 variants (19-47%) were expected to induce exon skipping by affecting ESRs.

We then performed an *ex vivo* splicing assay with pCAS2-*MLH1*e7+8 derived minigenes to experimentally verify these predictions (Figure 3). Surprisingly, the minigene assay revealed that only 1 out of the 32 variants (3%) altered the splicing pattern of *MLH1* exon 7. Moreover, this

variant (c.581T>G) did not lead to exon 7 skipping but to deletion of the last 8 nucleotides of the exon in about half of the minigene's transcripts (49%, Figure 3B). Further bioinformatics analysis using splice site-dedicated algorithms revealed that c.581T>G creates a 5'ss between exonic positions c.580 and c.581 (Figure S2D), which explains the observed 8-nucleotide deletion. Of note, our experimental results are concordant with those obtained for c.554T>G , which was shown not to disrupt splicing in patient RNA (Kohonen-Corish *et al.*, 1996), as well as for c.572G>T, c.577T>C, 578C>G and c.586A>T recently described to not alter exon 7 splicing by using a high-throughput reporter assay (Rhine *et al.*, 2018) (Table 1). Of note, out of these 5 variants, only c.586A>T, situated 3 nucleotides before the end of the exon, had been predicted to induce exon skipping both by MES+SSFL and the SRE-dedicated methods (Figure S2D & Table S4). Even though the number of true calls exceed that of false calls, and no negative false calls were produced, we observed that the SRE-dedicated bioinformatics tools generated a large number of false positive calls when predicting variant-increased exon skipping in the case of *MLH1* exon 7 (23-48%, depending on the tools taken into account, HAL being the SRE method showing the highest specificity and accuracy (Figure 3C & Table S4). Combining the predictions produced by the different SRE approaches did not significantly reduce the number of false calls relative to HAL as most of these were also produced by the other tools.

Overall, these results indicate that *MLH1* exon 7 seems resistant to ESR-mutations and that the newly developed SRE-dedicated *in silico* tools fail to correctly predict the impact on splicing of several variants mapping to this exon, suggesting that *MLH1* exon 7 could be a good model to better understand the limits of such tools.

Particular features of the exon 7 of *MLH1* may explain its resistance to exonic splicing mutations.

Next, we wondered if *MLH1* exon 7 could have particular characteristics conferring a natural resistance to mutations that affect potential ESR. To answer this question, we began by performing a bioinformatics comparative analysis of splice site strength between exon 7 and all other exons of *MLH1*. As shown on Figure 4A, our results revealed that the 5'ss and the 3'ss of exon 7 have very high scores (MES = 9.7 and SSFL = 87.6, and MES = 13.6 and SSFL = 99.6, respectively),

especially the 3' splice site that is predicted to be the strongest acceptor site of *MLH1*. Moreover, comparison of the relative frequencies of *MLH1* 9/6mer motifs at donor/acceptor sites showed that those of exon 7 (0.664% and 3.381%, respectively) have the highest values as compared to those of all other *MLH1* exons (Figure 4B). Interestingly, according to SROOGLE, the 5'ss of *MLH1* exon 7 has a MES score percentile of 0.71 relative to the scores of a dataset of over 50 000 constitutively spliced exons (i.e. only 29% of the exons within this dataset have higher 5'ss MES scores than *MLH1* exon 7), whereas the 3'ss has a score percentile of 1.0, which further underscores its exceptional predicted strength (Figure 4C). Finally, exon 7 has the particularity of being the smallest exon of *MLH1* with only 43 nucleotides and of being separated from the following exon (exon 8) by the shortest intron of this gene (143 nt) (Figure 4D). We suspected that the particularly important strengths of the 3'ss and/or of the 5'ss of *MLH1* exon 7 could protect the exon from potential ESR-mutations and explain the lack of reliability of the SRE-dedicated *in silico* predictions.

Donor and acceptor site strength confer resistance to alterations in potential exonic splicing regulatory elements of MLH1 exon 7.

In order to determine the contribution of *MLH1* exon 7 splice sites to the apparent resistance of this exon to potential ESR-mutations, we analyzed the effect of the 32 exonic variants in the context of modified pCAS2-*MLH1*.e7+8 minigenes carrying a suboptimal 5'ss or 3'ss at the level of exon 7. As we did not identify natural intronic splice site variants causing partial exon skipping (Figure 1B & Table 1), we decided to design suboptimal splice sites by first performing a bioinformatics analysis based on the MES and SSFL algorithms on a large number of hypothetical mutations mapping to intronic portions of *MLH1* exon 7 splice site consensus sequences (data not shown). We then pre-selected 8 of these variants for testing in the minigene reporter assay: 6 variants at the 5'ss and 2 at the 3'ss (Figure S3A). As shown in Figure S3B, these variants were predicted to decrease splice site strength to levels intermediate of those obtained with previously tested natural intronic variants, here used as reference points. The minigene assay showed that 6 out of the 8 variants indeed induce exon 7 skipping to different extents (from partial to total skipping depending on the variant) for the most part in agreement with *in silico* predictions (Figure S3C). Finally, we selected 2 leaky variants as the suboptimal 5'ss and 3'ss: c.588+6T>G ("+6TG context", 15% exon

skipping, ΔMES=-32%, ΔSSFL=-5.8) and c.546-3C>G ("-3CG context", ~37% exon skipping, ΔMES=-54%, ΔSSFL=-10.9), respectively. As determined by capillary electrophoresis, besides causing exon 7 skipping, the latter also leads to the activation of an extremely weak cryptic 3'ss between exonic positions c.553 and c.554 ($C_{551}A_{552}G_{553[}T_{554}$) causing deletion of the first 8 nucleotides of the exon in 13% of the minigene's transcripts (Δ7p(8nt), Figure S3). Curiously, this cryptic 3'ss is not predicted by SSFL and poorly predicted by MES both in the WT and mutated context, Figure S2E).

Next, we tested the impact of all natural exonic variants in both suboptimal splice site contexts. As shown in Figure 5B, we observed that in the context of a suboptimal 5'ss, 28 out of the 32 variants modified the splicing pattern of *MLH1* exon 7 when compared to the "+6TG context" alone, including 12 variations that increased exon skipping (38%), 15 that increased exon 7 inclusion (47%), and one (c.581T>G) that led to total deletion of the last 8 nucleotides of the exon. Only 4 variations (13%) had no impact on splicing. These data indicate that 5'ss strength explains, at least in part, the natural resistance of *MLH1* exon 7 to potential ESR-mutations. Indeed, except in 2 cases (c.586A>T and c.587A>C, which map to positions -3 and -2 of exon 7), all other variants causing exon skipping are located outside the sequences that define the splice sites. It is thus very likely that they affect splicing regulatory elements. As for c.586A>T and c.587A>C, although an eventual overlap between the 5'ss and ESRs cannot be excluded, it is probable that the observed defects are due to a direct decrease in 5'ss strength, as predicted by the MES and SSFL splice-site dedicated tools (Figure S2D).

Then, we analyzed the impact of the same series of 32 exonic variants in the context of the suboptimal 3'ss ("-3T>G context"). The minigene assay revealed that 28 out of the 32 variants (88%) altered *MLH1* exon 7 splicing when compared to c.548-3T>G alone, in part either by increasing exon skipping (n=12, 38%), increasing exon inclusion (n=15, 47%), or having complex effects (c.581T>G, 3%). whereas the 4 remaining variations (13%) had no impact on splicing (Figure 5C). As far as exon skipping is concerned, the results mostly mirrored those obtained in the "+6T>G context". We concluded therefore that 3'ss strength also contributes to the natural resistance of *MLH1* exon 7 to exonic variants potentially affecting SREs. Interestingly, most of the variations that increased exon inclusion in the context of the suboptimal 3'ss, including those located at a distance from the cryptic site (e.g. c.557A>G and c.568A>G), also increased,

proportionally and to a certain extent, the usage of the $C_{551}A_{552}G_{553}[T_{554}$ cryptic 3'ss (Figures 5C, S4) suggesting the implication of ESR in the recognition of both the natural and the cyptic 3'ss of exon 7. Still, the usage of the cryptic site was more affected by variants directly implicated in its definition, namely c.551C>A, c.551C>T, c.552A>T, c.553G>C, c.554T>A and c.554T>G (Figure S2E), which produced effects consistent with the known consensus features of 3'ss (Cartegni *et al.*, 2002).

In sum, apart from the usage of the cryptic 3'ss the minigene data obtained in the context of the suboptimal 3'ss are essentially equivalent to those obtained in the context of the suboptimal 5'ss. These results highlight an almost perfect symmetry between the contributions of 5'ss and 3'ss to the definition of exon 7, which is consistent with a cross-talk between the two splice sites over the exon, and indicate that both 5'ss and 3'ss contribute to the natural resistance of *MLH1* exon 7 to nucleotide variants potentially affecting ESR.

Suboptimal *MLH1* exon 7 splice sites restore the predictive power of SRE-dedicated *in silico* tools.

Given the variant-induced splicing defects detected in the contexts of suboptimal 5' and 3'ss, we decided to re-evaluate the performances of the SRE-dedicated tools in these conditions in particular for predicting variant-increased exon skipping. For this purpose, we separated the exonic variants into three groups according to the results obtained in the minigene assay as follows: (i) variants that increased exon 7 skipping (Δ7, n = 12), (ii) variants with no effect on exon 7 splicing (n = 4) and (iii) variants that increase exon 7 inclusion (either increasing FL, Δ7p(8nt), Δ7q(8nt) or Δ7p/q(8nt), n = 16) (Table S5).We then compared the experimental data with the bioinformatics predictions produced by QUEPASA, HEXplorer, SPANR, HAL and LR_{skip}, as previously described (Tubeuf et al, in preparation). Our results indicate that in these conditions the SRE-dedicated approaches can significantly discriminate which variants induce exon skipping from those that do not (t-test or its derivative, p-values <0.01) as well as predict the severity of the observed splicing defect (Figures S5 and S6, Pearson or Spearman correlation, p-values <0.001). Yet, none of the approaches was able to correctly separate the variants into three groups and predict the direction of the induced splicing defects (Figure S7, ANOVA's post-test or its derivative, p-

value ns) possibly, at least in part, due to a lack of statistical power (only four variants did not impair exon 7 splicing).

To better evaluate the performance of the SRE-dedicated approaches in predicting skipping of *MLH1* exon 7 bearing suboptimal splice sites, we next estimated the relative number of positive/negative and true/false calls produced by each approach according to previously described thresholds as indicated in Table S5, and then calculated the accuracy, specificity and sensibility of each method. As shown in Figure 5D, all the SRE-dedicated approaches displayed good predictive power in these conditions (accuracy ≥ 77%), with QUEPASA showing the best accuracy (88%) as well as good sensitivity (83%) and very good specificity (90%) (Figure 5D & Table S5).

Given the false calls discrepancies produced by the 4 SRE-dedicated approaches (Figure 5E), we then tested 3 multi-approach combinations to verify if the number of false calls could be reduced by taking advantage of the potentially complementary features of the different methods, similarly to what we recently described (Tubeuf *et al.*, in preparation). Here we found that LR_{skip} and AT LEAST 3 produced the lowest number of false calls (FC=3) outperforming QUEPASA&HAL (FC= 6) also in terms of sensitivity and accuracy (Figure 5D).We surmised that LRskip and AT LEAST 3 (91% and 91% accuracy) are more powerful than QUEPASA approach alone (91% accuracy), but that LR_{skip} has a slightly better sensitivity in this context. Altogether, our results demonstrate that the particularly important intrinsic strength of 5'ss and 3'ss can explain the unexpected number of false calls produced by the SRE-dedicated bioinformatics approaches in predicting variant-increased exon skipping in *MLH1* exon 7.

Discussion

MLH1, one of the major genes implicated in Lynch syndrome, was recently reported as being particularly affected by nucleotide variants that either alter splice sites or modify splicing regulatory elements (Rhine *et al.*, 2018). Part of these conclusions derived from the functional analysis of a relatively small number of variants (n=36) distributed in 5 *MLH1* exons, including exons 4, 5 and 7 for which no splicing defects were observed, and exons 8 and 15 in which an important fraction of variants caused exon skipping (6/6, and 5/7, respectively). These observations

led the authors to hypothesize that certain *MLH1* exons were more prone to splicing mutations than others. Here, we performed a thorough analysis of variants mapping in or near *MLH1* exon 7 (for which only 4 variants had been tested in the previous study) in order to better assess the vulnerability of this exon to splicing mutations. Our work focused on 63 variants: 31 intronic and 32 exonic. Concomitantly, we pursued the evaluation of the predictive power of splicing-dedicated *in silico* tools particularly focused on splice sites, branch point or SREs.

Before this study, only 6 intronic variants had been described experimentally as causing *MLH1* exon 7 aberrant splicing by disrupting the 3'ss or the 5'ss (Pagenstecher *et al.*, 2006; Planck *et al.*, 1999; Thompson *et al.*, 2013; Tournier *et al.*, 2008; Zavodna *et al.*, 2006; Pinero *et al.*, in preparation). Our work not only confirmed those initial findings but, importantly, uncovered 8 new splicing mutations, bringing the number of intronic variants negatively impacting *MLH1* exon 7 splicing to a total of 14. Our results indicated that most of the variations directly affecting the most conserved positions of the 3' and 5' splice sites have a drastic effect on exon 7 splicing, which was expected given the importance of these positions in splice site definition (Baralle *et al.*, 2009; Krawczak *et al.*, 2007; Leman *et al.*, 2018). Taken together, our data confirm the robustness of filtration strategies based on MES+SSFL and SPiCE prediction tools for pinpointing intronic variants potentially altering 3' or 5'splice sites (Houdayer *et al.*, 2012; Leman *et al.*, 2018; Spurdle *et al.*, 2008; Théry *et al.*, 2011) and encourage their use in molecular diagnostics as a decision-making tool to guide geneticists towards functional assessing the impact at the RNA level of potentially pathogenic variants. Importantly, we identified an intronic variant (c.588+38G>A) in *MLH1* intron 7 causing partial exon 7 skipping despite being located at a distance from the splice sites and thus escaping splice-site-dedicated predictions. We suspect that c.588+38G>A. either destroys an ISE or creates an ISS, or otherwise affects a potential RNA secondary structure hypothetically influencing the recognition of the 5'ss similarly to what it was described for *MAPT* exon 10 (Liu and Gong, 2008). To our knowledge, we thus identified the first variant in a MMR gene that induces a splicing defect by potentially affecting an intronic splicing regulatory element (ISR).

In contrast to the analysis of the intronic variants, we did not identify any exonic variant that could induce skipping of *MLH1* exon 7 by affecting potential ESRs. This was unexpected considering both the predictions of the recently developed SRE-dedicated *in silico* tools as well as the striking

high proportion of variant-induced ESR alterations detected in other exons including *MLH1* exon 10 (66%), *BRCA2* exons 7 and 18 (36% and 64%), *CFTR* exons 9 and 12 (77% and 54%), *FAS* exon 6 (61%), *SMN2* exon 7 (79%) and *WT1* exon 5 (64%) among others (Tubeuf *et al.*, in preparation, and references therein). Our results confirm the hypothesis from Rhine *et al.* that *MLH1* exons are not equally sensitive to exonic splicing mutations, exon 7 seeming particularly resistant. An explanation could be that *MLH1* exon 7 does not contain ESRs within its sequence or that it presents particular features that confer resistance to ESR-mutations. Both explanations offering as well a rationale for the unexpected high number of false positive calls produced by the SRE-dedicated tools in predicting variant-induced skipping of this exon.

Our computational analyses indicated that the strengths of both the 3'ss and the 5'ss of *MLH1* exon 7 seemed particularly important when compared to other exons both from *MLH1*and from other genes. Reducing the strength of each splice site by a mutagenesis approach revealed that the majority of the exonic variations that had no impact on splicing in the natural exon 7 context, induced splicing alterations in the suboptimal 3'ss and 5'ss contexts (27/31 variations, 87%). Thus, besides transforming *MLH1* exon 7 from a constitutively- into an alternatively- spliced exon, the suboptimal splice sites also rendered this exon very sensitive to ESR changes. Moreover, they restored the predictive power of the SRE-dedicated *in silico* approaches. These findings suggest that although *MLH1* exon 7 is not naturally sensitive to ESR-mutations, it does contain a high density of regulatory elements. The intrinsic strength of both splice sites appears to be sufficient, in the natural context, to not dependent on the auxiliary regulatory elements present along the exon (Fairbrother *et al.*, 2002). We cannot exclude however that alterations in these elements can have an effect on the splicing pattern of nearby exons. Indeed, recent studies have shown that certain splicing regulatory factors can bind to constitutive exons and have a long-distance influence on the regulation of upstream alternative exons probably by a competition mechanism (Ghigna *et al.*, 2005; Han *et al.*, 2011). It would be interesting to inspect if *MLH1* exon 7 variants have an impact on the level of alternative splicing of exon 6. Our findings support previously published data demonstrating that exon definition is equally influenced by the strengths of the 3 and 5splice sites (Shepard *et al.*, 2011) and explain, at least in part, the high number of false calls produced by the SRE-dedicated *in silico* tools in predicting variant-induced *MLH1* exon 7 skipping. We hypothesize that this explanation may hold true for other genes/exons. Further studies are needed to fully understand the determinants of exon vulnerability to ESR-mutations, which may in turn allow to

define new parameters for improving current SRE-dedicated bioinformatics approaches. It is conceivable that the integration of exon-specific features in the calculations/interpretations of SRE-dedicated approaches, such as the combined 3'ss and 5'ss strengths (Shepard *et al.*, 2011) will in the future contribute to improve the current algorithms.

Most of the variants analyzed in this study were identified in patients suspected of Lynch syndrome but remained classified as VUS and, as such, are non-actionable i.e. cannot be used in clinical decisions. Our study not only contributed to better characterize *MLH1* exon 7 splicing signals but it also provided functional evidence that may have important implications for the molecular diagnosis of Lynch syndrome. Skipping of *MLH1* exon 7 leads to a frameshift, resulting in the introduction of a premature termination codon in exon 8 probably targeting the aberrant *MLH1* transcripts to degradation by the nonsense mediated mRNA decay, or otherwise causing the production of a MLH1 protein carrying a C-terminal deletion (p.Arg182Serfs*6). Consequently, variants inducing total skipping of exon 7 or other truncating changes in this exon (including indels or nonsense variants not affecting splicing) cause a drastic loss in full-length MLH1 protein and can in principle be considered deleterious. This is the classification we now suggest for 21 out of the 63 variants analyzed in this study (Table 1). Conversely, the 15 intronic variants and 7 exonic synonymous substitutions that had no impact on splicing in our minigene assay are likely to be non-pathogenic, whereas those inducing partial splicing defects (c.581T>G and c.588+38G>A) should still be considered as VUS (Table 1) given that the minimal amount of full-length transcripts necessary to fulfil *MLH1* function is currently unknown. In any case, the final clinical classification of many of these variants may require further studies, including quantitative analyses of patient's RNA, protein assays for missense variants, as well as an evaluation of patient clinical history and family data, and their global assessment by expert panels (Thompson *et al.*, 2014).

Our work on *MLH1* exon 7 contributes to the characterization the cis-elements that determine the efficient splicing of this constitutive exon, to a better understanding of the advantages and limitations of splicing-dedicated *in silico* tools, and to the biological assessment of unclassified Lynch syndrome-associated variants. Altogether, our data should contribute, in the future, to improve splicing-based bioinformatics analysis in particular those aiming at pinpointing SRE alterations, the final goal being to prioritize disease-causing candidates for functional analysis among the large number of VUS currently detected by molecular diagnostic laboratories. .

Bibliography

Ars, E., Serra, E., García, J., Kruyer, H., Gaona, A., Lázaro, C., and Estivill, X. (2000). Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum. Mol. Genet. *9*, 237–247.

Auclair, J., Busine, M.P., Navarro, C., Ruano, E., Montmain, G., Desseigne, F., Saurin, J.C., Lasset, C., Bonadona, V., Giraud, S., *et al.* (2006). Systematic mRNA analysis for the effect of MLH1 and MSH2 missense and silent mutations on aberrant splicing. Hum. Mutat. *27*, 145–154.

Baralle, D., and Buratti, E. (2017). RNA splicing in human disease and in the clinic. Clin. Sci. Lond. Engl. 1979 *131*, 355–368.

Baralle, D., Lucassen, A., and Buratti, E. (2009). Missed threads. The impact of pre-mRNA splicing defects on clinical practice. EMBO Rep. *10*, 810–816.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. BMC Bioinformatics *15*, 293.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., and Martins, A. (2013). 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. *34*, 1547–1557.

Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nat. Rev. Genet. *18*, 599–612.

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J.O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res. *42*, 10681–10697.

Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007–1013.

Frebourg, T. (2014). The challenge for the next generation of medical geneticists. Hum. Mutat. *35*, 909–911.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. Clifton NJ *653*, 249–257.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res. *36*, 2257–2267.

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P.M., Green, M.R., Riva, S., and Biamonti, G. (2005). Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. Mol. Cell *20*, 881–890.

Grandval, P., Fabre, A.J., Gaildrat, P., Baert-Desurmont, S., Buisine, M.-P., Ferrari, A., Wang, Q., Béroud, C., and Olschwang, S. (2013). UMD-MLH1/MSH2/MSH6 databases: description and analysis of genetic variations in French Lynch syndrome families. Database J. Biol. Databases Curation *2013*, bat036.

Grodecká, L., Buratti, E., and Freiberger, T. (2017). Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? Int. J. Mol. Sci. *18*.

Han, J., Ding, J.-H., Byeon, C.W., Kim, J.H., Hertel, K.J., Jeong, S., and Fu, X.-D. (2011). SR proteins induce alternative exon skipping through their activities on the flanking constitutive exons. Mol. Cell. Biol. *31*, 793–802.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., and Pease, L.R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene *77*, 51–59.

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., *et al.* (2012). 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. *33*, 1228–1238.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

van der Klift, H.M., Jansen, A.M.L., van der Steenstraten, N., Bik, E.C., Tops, C.M.J., Devilee, P., and Wijnen, J.T. (2015). 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. *3*, 327–345.

Kohonen-Corish, M., Ross, V.L., Doe, W.F., Kool, D.A., Edkins, E., Faragher, I., Wijnen, J., Khan, P.M., Macrae, F., and St John, D.J. (1996). RNA-based mutation screening in hereditary nonpolyposis colorectal cancer. Am. J. Hum. Genet. *59*, 818–824.

Kol, G., Lev-Maor, G., and Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. Hum. Mol. Genet. *14*, 1559–1568.

Královicová, J., Lei, H., and Vorechovský, I. (2006). Phenotypic consequences of branch point substitutions. Hum. Mutat. *27*, 803–813.

Krawczak, M., Thomas, N.S.T., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., and Cooper, D.N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum. Mutat. *28*, 150–158.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980-985.

Lastella, P., Surdo, N.C., Resta, N., Guanti, G., and Stella, A. (2006). *In silico* and in vivo splicing analysis of MLH1 and MSH2 missense mutations shows exon- and tissue-specific effects. BMC Genomics *7*, 243.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Leman, R., Gaildrat, P., Gac, G.L., Ka, C., Fichou, Y., Audrezet, M.-P., Caux-Moncoutier, V., Caputo, S.M., Boutry-Kryza, N., Léone, M., *et al.* (2018). 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.

Lewandowska, M.A. (2013). The missing puzzle piece: splicing mutations. Int. J. Clin. Exp. Pathol. *6*, 2675–2682.

Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J., and Fairbrother, W.G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc. Natl. Acad. Sci. U. S. A. *108*, 11093–11098.

Liu, F., and Gong, C.-X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol. Neurodegener. *3*, 8.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. *579*, 1900–1903.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res. *25*, 290–303.

Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., Santamariña, M., Behar, R., Blanco, A., Carrasco, E., *et al.* (2018). Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? Front. Genet. *9*, 366.

Mort, M., Sterne-Weiler, T., Li, B., Ball, E.V., Cooper, D.N., Radivojac, P., Sanford, J.R., and Mooney, S.D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol. *15*, R19.

Motohashi, K. (2015). A simple and efficient seamless DNA cloning method using SLiCE from Escherichia coli laboratory strains and its application to SLiP site-directed mutagenesis. BMC Biotechnol. *15*, 47.

Pagenstecher, C., Wehner, M., Friedl, W., Rahner, N., Aretz, S., Friedrichs, N., Sengteller, M., Henn, W., Buettner, R., Propping, P., *et al.* (2006). Aberrant splicing in MLH1 and MSH2 due to exonic and intronic variants. Hum. Genet. *119*, 9–22.

Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. *32*, 577–591.

Planck, M., Koul, A., Fernebro, E., Borg, A., Kristoffersson, U., Olsson, H., Wenngren, E., Mangell, P., and Nilbert, M. (1999). hMLH1, hMSH2 and hMSH6 mutations in hereditary non-polyposis colorectal cancer families from southern Sweden. Int. J. Cancer *83*, 197–202.

Plazzer, J.P., Sijmons, R.H., Woods, M.O., Peltomäki, P., Thompson, B., Den Dunnen, J.T., and Macrae, F. (2013). The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. Fam. Cancer *12*, 175–180.

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. J. Hum. Genet. *59*, 5–15.

Rhine, C.L., Cygan, K.J., Soemedi, R., Maguire, S., Murray, M.F., Monaghan, S.F., and Fairbrother, W.G. (2018). Hereditary cancer genes are highly susceptible to splicing mutations. PLoS Genet. *14*, e1007231.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., *et al.* (2015). 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. *17*, 405–424.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Schwartz, S., Hall, E., and Ast, G. (2009). SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. Nucleic Acids Res. *37*, W189-192.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. *18*, 88–103.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155–7174.

Shendure, J. (2011). Next-generation human genetics. Genome Biol. *12*, 408.

Shepard, P.J., Choi, E.-A., Busch, A., and Hertel, K.J. (2011). Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites. Nucleic Acids Res. *39*, 8928–8937.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., Tosi, M., and Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using *In silico* Tools. PLoS Genet. *12*, e1005756.

Spurdle, A.B., Couch, F.J., Hogervorst, F.B.L., Radice, P., Sinilnikova, O.M., and IARC Unclassified Genetic Variants Working Group (2008). Prediction and assessment of splicing alterations: implications for clinical testing. Hum. Mutat. *29*, 1304–1313.

Teraoka, S.N., Telatar, M., Becker-Catania, S., Liang, T., Onengüt, S., Tolun, A., Chessa, L., Sanal, O., Bernatowska, E., Gatti, R.A., *et al.* (1999). Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am. J. Hum. Genet. *64*, 1617–1631.

Théry, J.C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M.-P., Killian, A., Duponchel, C., Rousselin, A., Vaur, D., Peyrat, J.-P., *et al.* (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur. J. Hum. Genet. *19*, 1052–1058.

Thompson, B.A., Goldgar, D.E., Paterson, C., Clendenning, M., Walters, R., Arnold, S., Parsons, M.T., Michael D, W., Gallinger, S., Haile, R.W., *et al.* (2013). A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. Hum. Mutat. *34*, 200–209.

Thompson, B.A., Spurdle, A.B., Plazzer, J.-P., Greenblatt, M.S., Akagi, K., Al-Mulla, F., Bapat, B., Bernstein, I., Capellá, G., den Dunnen, J.T., *et al.* (2014). Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat. Genet. *46*, 107–115.

Thompson, B.A., Martins, A., and Spurdle, A.B. (2015). A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays. Clin. Genet. *87*, 100–108.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J., *et al.* (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. *29*, 1412–1424.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA N. Y. N *14*, 802–813.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. J. Comput. Mol. Cell Biol. *11*, 377–394.

Zavodna, K., Bujalkova, M., Krivulcik, T., Alemayehu, A., Skorvaga, M., Marra, G., Fridrichova, I., Jiricny, J., and Bartosova, Z. (2006). Novel and recurrent germline alterations in the MLH1 and MSH2 genes identified in hereditary nonpolyposis colorectal cancer patients in Slovakia. Neoplasma *53*, 269–276.

Legends to figures

Table 1. Naturally occuring *MLH1* exon 7 variants selected for this study. Nucleotide variants in or near *MLH1* exon 7 were collected from 8 human variation databases (COSMIC, ClinVar, dbSNP, ESP, gnomAD, HGMD, LOVD and UMD) by interrogating the DNA sequence spanning MLH1 positions c.546-40 to c.588+40.

[1] Variant classification was retrieved from each database when available and refers to the 5-tier system used by the InSiGHT Variant Interpretation Committee (http://insight-group.org/variants/classifications/) as follows: 1, not pathogenic; 2, likely not pathogenic; 3, uncertain significance (also called VUS for variants of unknown significance); 4, likely pathogenic; 5, pathogenic. n/a, not available.

[2] Previous reports derived from cell-based minigene assays (Rhine *et al.*, 2018[1]; Tournier *et al.*, 2011[2]; Pinero *et al.*, in preparation[3]).

[3] Previously described data derived from RT-PCR analyses of patients' RNA (Zavodna *et al.*, 2016[1]; Planck *et al.*, 1999[2], Kohonen-Corish *et al.*, 1996[3]; Thompson *et al.*, 2013[4]; Pagenstecher *et al.* 2006[5], Farrington *et al.*, 1998[6]).

Figure 1. A relatively important fraction of intronic variants alters the splicing pattern of *MLH1* exon 7 in agreement with *in silico* predictions. (A) Distribution of the 31 natural intronic variants reported in the vincity of *MLH1* exon 7. The diagram shows the nucleotide composition of *MLH1* exon 7 and its flanking intronic regions (c.546-40_c.546-40), the amino-acid sequence

encoded by exon 7 (1-letter code, p.182_p.196), as well as the relative position and identity of each variant. (B) RT-PCR analysis of the splicing pattern of pCAS2-*MLH1*e7-8 minigenes carrying the variants of interest. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products are indicated on the right. (C) Comparison of the performances of splice site-dedicated bioinformatics approaches in predicting variant-induced alterations in *MLH1* exon 7 3'ss and 5'ss. Results are based on the comparison of the experimental data obtained in (B) with the *in silico* evaluation presented in Table S3. *In silico* predictions of potential splice site alterations were conducted by using MES and SSFL as well as MES+SSFL and SPiCE, as described under Materials and Methods. True and false calls of variant-induced exon-skipping were determined by taking into account the following thresholds: -15% for MES, -5% for SSFL (as suggested in Houdayer *et al.*, 2012) and 11,5% for SPiCE (as suggested in Leman *et al.*, 2018).

Figure 2. The adenosine at position -23 serves as branch point in *MLH1* intron 6, as predicted by bioinformatics approaches, but is dispensable for efficient splicing of exon 7. (A) Identification by *in silico* predictions tools of potential *MLH1* exon 7 branch points (BPs). The diagram shows the nucleotide composition of *MLH1* intron 6 (c.546-100-c.546-1) as well as the relative position and scores of the putative BPs predicted by HSF, SSFL and SROOGLE (a, scores based on Kol *et al.*, 2005 and b, scores based on Schwartz *et al.*, 2008). (B) *In silico* predictions by HSF, SSFL and SROOGLE of variant-induced alterations of putative *MLH1* exon 7 branch points. (C) RT-PCR analysis of the splicing pattern of pCAS2-*MLH1*e7-8 minigenes carrying artificial variants replacing the putative BPs. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products obtained are indicated on the right. (D) Strategy to experimentally map the branch point of *MLH1* exon 7. The image represents the lariat protected from RNase R digestion at the 5'ss-BP junction. Arrows indicate the primers used during the first and the second step of the nested RT-PCR reactions performed on endogenous transcripts of a human cell line (F1 and R1, and F2 and R2, respectively)

as described under Materials and Methods. (E) DNA sequencing electropherogram of the nested RT-PCR product corresponding to the *MLH1* intron 6 lariat showing the exact position of the branch point (BP, black triangle). Sequencing was performed with the F2 primer.

Figure 3. Contrary to SRE-dedicated *in silico* predictions none of the variants mapping to *MLH1* exon 7 induce exon skipping. (A) Distribution of the 32 natural variants reported within *MLH1* exon 7. The diagram shows the nucleotide composition of *MLH1* exon 7, the amino-acid sequence encoded by exon 7 (1-letter code, p.182_p.196), as well as the relative position and identity of each variant. (B) RT-PCR analysis of the splicing pattern of pCAS2-*MLH1*e7-8 minigenes carrying the exonic variants of interest. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products are indicated on the right. The splicing pattern of c.588del and c.588dup was not altered (100% exon inclusion) but the included exon 7 either lacked its last nucleotide or contained a 1-nt duplication as specified by each mutation. (C) Summary of the performances of SRE-dedicated bioinformatics approaches (QUEPASA, HEXplorer, SPANR and HAL as well as QUEPASA&HAL, AT LEAST 3 and LRskip) in predicting variant-induced *MLH1* exon 7 skipping. Results are based on the comparison of the experimental data described in (B) with the *in silico* predictions presented in Table S4. True and false calls were determined by taking into account the following thresholds as previously described (Tubeuf et al, in preparation): -0.5 for QUEPASA, -14 for HEXplorer, -0.1% for SPANR, -3.4% for HAL and 31.1% for LRskip. (D) Comparison, by using a Venn diagram, of the false calls produced by QUEPASA, HEXplorer, SPANR and HAL in predicting variant-induced skipping of *MLH1* exon 7.

Figure 4. Features of *MLH1* exons. (A) Strength of natural *MLH1* donor (5'ss) and acceptor (3'ss) sites as predicted by MES and SSFL splice site-dedicated *in silico* tools. (B) Frequency of *MLH1* donor (9 mer) and acceptor (6 mer) splice sites relative to 10728 human donor and acceptor sites as provided by the Alamut Visual interface. (C) MES score percentiles of natural MLH1 5'ss and a3'ss s relative to the MES scores of a dataset of over 50 000 constitutively or alternatively spliced exons as provided by SROOGLE. (D) Size of *MLH1* exons and their flanking introns.

Figure 5. Weakening the 5'ss or the 3'ss renders *MLH1* exon 7 sensitive to ESR mutations and unveils the predictive power of SRE-dedicated *in silico* tools. (A) Distribution of the 32 natural variants reported within *MLH1* exon 7 and of the 2 artificial intronic variants designed to either weaken the 5'ss (c.588+6T>G) or the 3'ss (c.546-3C>G). The diagram shows the nucleotide composition of *MLH1* exon 7 and part of the flanking intronic sequences, the amino-acid sequence encoded by exon 7 (1-letter code, p.182_p.196), as well as the relative position and identity of each variant. (B) RT-PCR analysis of the splicing pattern of pCAS2-*MLH1*e7-8 c.588+6T>G minigenes (i.e. with a suboptimal 5'ss) carrying the exonic variants of interest. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products obtained are indicated on the right. "E7 inclusion Δ8nt start&end" indicates the concomitant deletion of the first and the last 8 nucleotides of the exon, whereas "E7 inclusion Δ8nt start/end" corresponds to RT-PCR products of identical size (deleted of the first or the last 8 nucleotides of the exon) that were detected by Sanger sequencing but cannot be discriminated by electrophoresis. (C) RT-PCR analysis of the splicing pattern of pCAS2-*MLH1*e7-8 c.546-3C>G minigenes (i.e. with suboptimal 3'ss) carrying the exonic variants of interest, similarly to what was described in (B). (D) Comparison of performances of SRE-dedicated bioinformatics approaches (QUEPASA, HEXplorer, SPANR and HAL as well as QUEPASA&HAL, AT LEAST 3 and LRskip) in predicting variant-induced skipping of *MLH1* exon 7 in the context of the suboptimal splice sites. Results are based on the comparison of the experimental data described in (B) and (C) with the *in silico* predictions presented in Table S5. True and false calls were determined by taking into account the following thresholds as previously described (Tubeuf et al, in preparation): -0.5 for QUEPASA, -14 for HEXplorer, -0.1% for SPANR, -3.4% for HAL and 31.1% for LRskip. (E) Comparison, by using a Venn diagram, of the false calls produced by QUEPASA, HEXplorer, SPANR and HAL in predicting exon skipping events in the context of the suboptimal splice sites.

Figure S1. Structure of the pCAS2-*MLH1*e7-8 minigene used in the cell-based splicing reporter assays. The pCAS2-*MLH1*e7-8 minigenes were generated by inserting a genomic fragment containing *MLH1* exons 7 and 8 as well as upstream/downstream flanking intronic sequences (157 and 231 nucleotides, respectively) into the intron of the pCAS2 vector, as indicated. The pCAS2 vector was described previously (Soukarieh *et al.*, 2016) and carries two exons (A and

B) with a sequence derived from the human *SERPING1/C1NH* gene, separated by an intron containing BamHI and MluI cloning sites. Boxes represent exons, and lines in between indicate introns, whereas the bent arrow specifies the cytomegalovirus (CMV) promoter and the black circle indicates the polyadenylation signal (Poly A). Arrows below the exons represent primers used in RT-PCR reactions. The star in the forward primer symbolizes a 6-FAM 5' fluorescent modification for detection of the RT-PCR products upon capillary electrophoresis.

Figure S2. Splice-site dedicated bioinformatics predictions of variant-induced creation of de novo splice sites or activation of cryptic splice sites experimentally detected in the minigene assays. Splice site-related *in silico* analyses were performed with MaxEntScan (MES), SpliceSiteFinder-like (SSFL) and/or SPiCE, as described under Materials and Methods. (A) According to MES and SSFL, *MLH1* c.574_588+2del could lead to the creation of a new 5'ss between positions c.571 and c.572 concomitant to the deletion of the natural 5'ss of exon 7. (B) *MLH1* c.546-2A>C, c.546-2A>G and c.546-2A>T were predicted to destroy the natural 3'ss and simultaneously create a new 3'ss between positions c.553 and c.554. *MLH1* c.546-1G>A was predicted to destroy the natural 3'ss and to create a weaker 3'ss one nucleotide downstream between positions c.546 and c.547. (C) *MLH1* c.581T>G does not affect the natural 5'ss but it was predicted to create a new 5'ss between c.580 and c.581. (D) The predicted impact on 5'ss strength of exonic variants overlapping the 5'ss sequence (c.586A>T, c.587A>C, c.588delA and c.588dup) was assessed both in the natural (WT) and in the artificial suboptimal 5'ss contexts (c.588+6T>G). MES and SSFL results are presented as the change in scores (ΔMES and ΔSSFL, respectively) of the variants relative to the corresponding reference sequence (WT or c.588+6T>G). Scores smaller than the indicated thresholds (-15% for ΔMES, -5% for ΔSSFL and 11.5% for SPiCE) were considered as predictive of splicing alterations. n/a., not applicable. (E) *MLH1* c.546-3C>G was predicted to decrease the strength of the natural 3'ss and not to significantly affect a very weak cryptic 3'ss, predicted by MES only, located between positions c.553 and c.554. This cryptic site was predicted to be damaged by c.551C>A, c.552A>T and c.553G>C, and reinforced by c.554T>G and c.554T>A.

Figure S3. Impact on splicing of artificial intronic variations predicted to decrease the strength of *MLH1* exon 7 splice sites. (A) Distribution of 12 intronic variations predicted to decrease the strength of *MLH1* exon 7 splice sites, including 4 natural variants already tested in the

minigene assay (Figure1), here used as controls, and 8 artificial variants (indicated by a dot) designed to gradually weaken the 3'ss and the 5'ss of exon 7. The diagram shows the nucleotide composition of *MLH1* exon 7 and part of its flanking introns, the amino-acid sequence encoded by the exon (1-letter code), as well as the relative position and identity of the variants of interest. (B) Variant-induced alterations in the strength of *MLH1* exon 7 splice sites as predicted by the MES and SSFL *in silico* tools (ΔMES and ΔSSFL values relative to WT). (C) RT-PCR analysis of the splicing pattern of pCAS2-*MLH1*e7-8 minigenes carrying the intronic variants of interest. The top panel shows the RT-PCR products separated on an agarose gel, whereas the bottom panel indicates the relative quantification of equivalent fluorescent RT-PCR products separated by capillary electrophoresis. Results represent the mean ± SEM of three independent transfection experiments. The identities of the RT-PCR products are indicated on the right.

Figure S4. Correlation between variant-associated exon skipping levels and cryptic 3'ss activation obtained in the context of a suboptimal 3'ss. Exon skipping (Δ7) levels and cryptic 3'ss activation (Δ7p(8nt)) refer to semi-quantitative data obtained from the pCAS2-*MLH1*e7-8 c.546-3C>G minigene assay. Determination coefficients (R^2) and two-sided p-value were determined by performing a Spearman correlation analysis. Based on bioinformatics analysis (Figure S2E), variants directly affecting the $C_{551}A_{552}G_{553}[T_{554}$ cryptic 3'ss (c.551C>A, c.552A>T, c.553G>C, c.554T>A and c.554T>G) were excluded from this analysis.

Figure S5. Correlation between variant-associated exon skipping levels described in the context of a suboptimal 5'ss and *in silico* data obtained with SRE-dedicated approaches. Exon skipping levels refer to semi-quantitative data obtained from the pCAS2-*MLH1*e7-8 c.588+6T>G minigene assay. Panels A to F compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL, LRskip and LRinc, respectively (Supplementary Table S5). Determination coefficients (R^2) and two-sided p-values were determined by performing a Pearson or Spearman correlation analysis, as indicated in Supplementary Table S2.

Figure S6. Correlation between variant-associated exon skipping levels described in the context of a suboptimal 3'ss and *in silico* data obtained with SRE-dedicated approaches. Exon skipping levels refer to semi-quantitative data obtained from the pCAS2-*MLH1*e7-8 c.546-3C>G minigene assay. Panels A to F compare these data with the corresponding *in silico* results obtained

with QUEPASA, HEXplorer, SPANR, HAL, LRskip and LRinc, respectively (Supplementary Table S5). Determination coefficients (R^2) and two-sided p-values were determined by performing a Pearson or Spearman correlation analysis, as indicated in Supplementary Table S2..

Figure S7. Comparison of the variant-associated splicing effects observed in the context of suboptimal splice sites with *in silico* data obtained with SRE-dedicated approaches. The 32 exonic variations were separated into 3 groups according to their impact on splicing as experimentally determined in Figure 5 and described in Supplementary Table S5. Panels A to F compare these data with the corresponding *in silico* results obtained with QUEPASA, HEXplorer, SPANR, HAL, LRskip and LRinc, respectively (Supplementary Table S5). The dashed lines indicate the thresholds used in this study as shown in Tables S4 and S5. Two-sided p-values were calculated by using ANOVA or Kruskal-Wallis, as indicated in Supplementary Table S2.

Table S1. Description of the primers used in this study.

[1] F, forward; R, reverse.

[2] The position of the MLH1 variations are underlined. The double underlined sequences correspond to BamHI and MluI restriction sites and the sequence highlighted in grey correspond to the 15nt-tail used for homologous recombination.

Table S2. Description of statistical analyses conducted in this study. Data derived from confrontation of experimental and *in silico* analyses were compared by using either Student's test, one-way ANOVA test and Pearson's correlation or their derivatives depending on the purpose of the analysis and data distribution patterns.

Table S3. Comparison of intronic variant-associated splicing effects observed in the pCAS2-*MLH1*e7-8 minigenes and associated splice site-dedicated bioinformatics approaches. The impact on splicing of 31 intronic variants located in or near the splice sites of *MLH1* exon 7 was determined by performing a cell-based splicing assay with pCAS2-*MLH1*e7-8 minigenes. Δ7, exon skipping; Δ7p(8nt) and Δ7q(17nt), deletion of the first 8/last 17 nt of exon 7. The table shows a separation of the variants into 2 groups according to the minigene results shown in Figure 1: variants that induced splicing defects (n=14) and those that did not (n=17). *In silico* predictions of

potential effects on splicing were conducted by using 2 splice site-dedicated *in silico* tools (MES, SSFL), as well as two approaches based on their combination (MES+SSFL and SPiCE). MES and SSFL results are presented as the change in scores (Δ) of the variants relative to WT (ΔMES and ΔSSFL, respectively). True and false calls (in grey) of exon 7 splicing defects were determined by taking into account the following thresholds: -15% for ΔMES, -5% for ΔSSFL and 11.5% for SPiCE as previously recommended (Houdayer *et al.*, 2012; Leman *et al.*, 2018) and indicated between parenthesis in the Table.

Table S4. Comparison of exonic variant-associated splicing effects observed in the pCAS2-MLH1e7-8 minigenes and associated SRE-dedicated *in silico* predictions. The impact on splicing of 32 variants mapping *MLH1* exon 7 was determined by performing a cell-based splicing assay with pCAS2-*MLH1*e7-8 minigenes as shown in Figure 3. These variants were retrieved from human variation databases and those overlapping to splice sites are indicated by a star. Δ7q(8nt), deletion of the last 8 nt of exon 7. *In silico* predictions of potential effects on ESRs were conducted by using the 4 new SRE-dedicated *in silico* tools (QUEPASA, HEXplorer, SPANR and HAL), as well as three approaches resulting from their combination (QUEPASA&HAL, at AT LEAST 3 and LRskip). True and false calls (in grey) for prediction of induced exon-skipping events were determined by taking into account the following thresholds as previously recommended (Tubeuf *et al.*, in preparation) and indicated between parenthesis in the Table: -0.50 for QUEPASA (ΔtESRseq scores), -14 for HEXplorer (ΔHZEI scores), -0.1% for SPANR (Δψ scores), -3.4% for HAL (Δψ scores) and 31.1% for LRskip. n/a, not applicable.

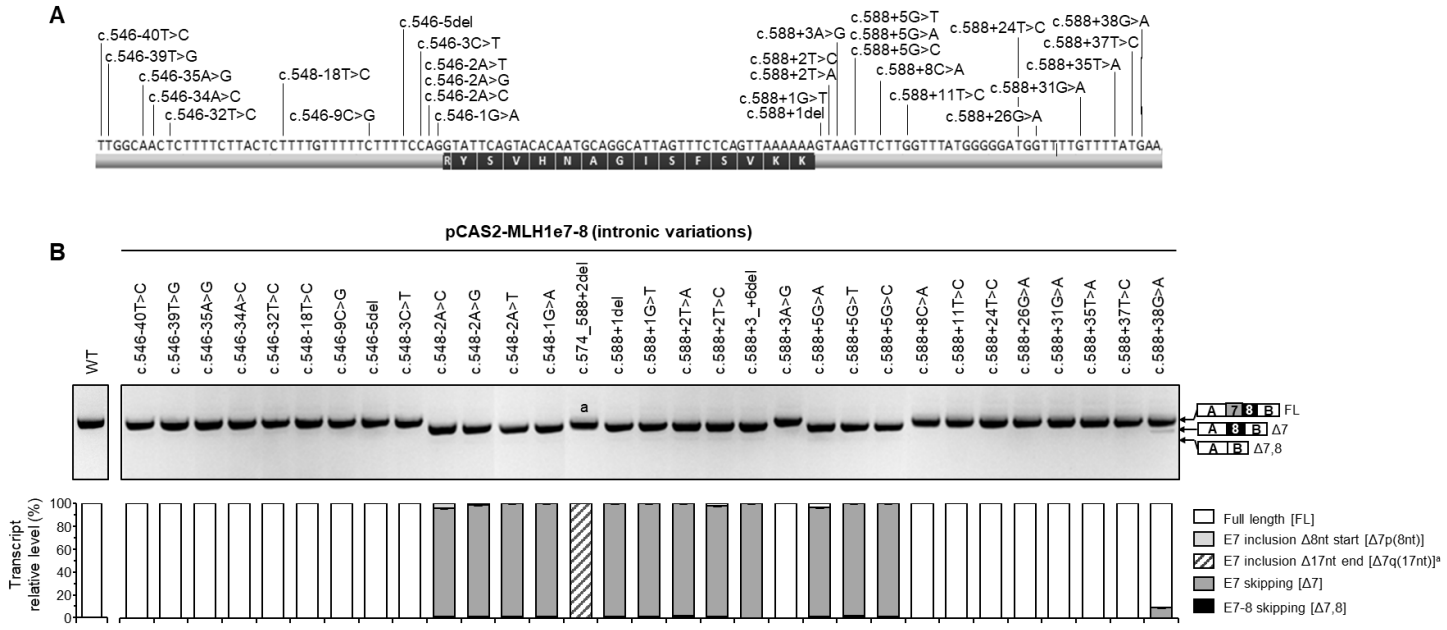
Table S5. Comparison of exonic variant-associated splicing effects observed in the context of suboptimal splice sites and associated SRE-dedicated *in silico* predictions. The impact on splicing of 32 variants mapping *MLH1* exon 7 was determined by performing a cell-based splicing assay with pCAS2-*MLH1*e7-8 minigenes also bearing a suboptimal 3' or 5'ss (c.546-3C>G or c.588+6T>G, respectively), as shown in Figure 3. These variants were retrieved from human variation databases and those overlapping to splice sites are indicated by a star. Δ7q(8nt), deletion of the last 8 nt of exon 7. Variants were separated into 3 groups according to their impact on splicing as experimentally determined in Figure 5: variants that increased exon 7 skipping (n=12), variations with no effect on exon 7 splicing (n = 4) and those that increased exon 7 inclusion (n = 16). Variation increasing either FL, (inclusion of exon 7), Δ7p(8nt) (inclusion of exon 7 deleted of the

first 8 nt), Δ7q(8nt) (inclusion of exon 7 deleted of the last 8 nt) or Δ7p(8nt)/q(8nt) (inclusion of exon 7 deleted of the first and the last 8 nt) were considered as increasing exon 7 inclusion. *In silico* predictions of potential effects on ESRs were conducted by using the 4 new SRE-dedicated *in silico* tools (QUEPASA, HEXplorer, SPANR and HAL), as well as three approaches resulting from their combination (QUEPASA&HAL, at AT LEAST 3 and LRskip). True and false calls (in grey) for prediction of induced exon-skipping events were determined by taking into account the following thresholds as previously recommended (Tubeuf *et al.*, in preparation) and indicated between parenthesis in the Table: -0.50 for QUEPASA (ΔtESRseq scores), -14 for HEXplorer (ΔHZEI scores), -0.1% for SPANR (Δψ scores), -3.4% for HAL (Δψ scores) and 31.1% for LRskip. n/a, not applicable.

Variations			Databases Clinical classification[1]								Splicing data		
Positions	Nucleotide variations	Predicted protein changes	COSMIC	ClinVar	dbSNP	ESP	gnmoAD	HGMD	LOVD	UMD	Minigene splicing assay[2]	Patient's RNA analysis[3]	Classification suggested
Intron 6 (n=13)	c.546-40T>C	p.?			n/a	n/a	n/a				No effect		1
	c.546-39T>G	p.?								2	No effect		1
	c.546-35A>G	p.?			n/a	n/a					No effect		1
	c.546-34A>C	p.?			n/a		n/a				No effect		1
	c.546-32T>C	p.?	n/a		n/a				n/a		No effect		1
	c.546-18T>C	p.?		2	2		n/a			3	No effect		1
	c.546-9C>G	p.?		2	2		n/a				No effect		1
	c.546-5del	p.?		1 | 3	2 | 3		n/a				No effect		1
	c.546-3C>T	p.?			n/a		n/a				No effect		1
	c.546-2A>C	p.?		5	5			5	5		Total effect (Δ7)	✓ [1]	5
	c.546-2A>G	p.?		5	5			5	4 | 5	5	Total effect (Δ7)	✓ [2]	5
	c.546-2A>T	p.?								5	Total effect (Δ7)		5
	c.546-1G>A	p.?	n/a	4 | 5	4			5	4		Total effect (Δ7)		5
Exon 7 (n=33)	c.551C>A	p.Ser184*	n/a					5	n/a		No effect		3
	c.551C>T	p.Ser184Leu		3							No effect		3
	c.552A>T	p.=		2 | 3	2		n/a				No effect		1
	c.553G>C	p.Val185Leu			n/a						No effect		3
	c.554T>A	p.Val185Glu		3	3 | 5						No effect		3
	c.554T>G	p.Val185Gly		4 | 5	3 | 5			5	5		No effect	✓ [3]	3
	c.556C>T	p.His186Tyr			n/a						No effect		3
	c.557A>C	p.His186Pro		3					n/a		No effect		3
	c.558C>T	p.=		2 | 3	3				3		No effect		1
	c.559A>T	p.Asn187Tyr	n/a								No effect		3
	c.563C>T	p.Ala188Val			n/a		n/a				No effect		3
	c.565G>A	p.Gly189Ser		3	3						No effect		3
	c.567C>G	p. =	n/a								No effect		1
	c.568A>G	p.Ile190Val		3	3		n/a				No effect		3
	c.568delA	p.Ile190Leufs*12						5			No effect		5
	c.572G>T	p.Ser191Ile		3	3			4	3		No effect [1]		3
	c.573T>C	p.=		2	n/a		n/a				No effect		1
	c.574_588+2del	p.(Phe192_ Lys196del)		4				5	4		Total effect [FLΔq(17nt)]		5
	c.576C>T	p.=	n/a	2							No effect		1
	c.577T>C	p.Ser193Pro		3	3			5	3		No effect [1]		3
	c.578C>A	p.Ser193*								5	No effect		5
	c.578C>G	p.Ser193*		5	5			5	5		No effect [1]		5
	c.578C>T	p.Ser193Leu			5						No effect		3
	c.579A>G	p.=	n/a	2 | 3	1 | 2		n/a			1	No effect		1
	c.580G>A	p.Val194Ile	n/a				n/a				No effect		3
	c.581T>G	p.Val194Gly		3	3						Partial effect [FLΔq(8nt)]		3
	c.582T>C	p.=								3	No effect		1
	c.582del	p.Lys196Asnfs*6	n/a					5	n/a		No effect		5
	c.583A>T	p.Lys195*	n/a	4 | 5	4 | 5			5			No effect		5
	c.586A>T	p.Lys196*		5	5			5	5	5	No effect [1]		5
	c.587A>C	p.Lys196Thr		3	3						No effect		3
	c.588delA	p.Lys196Asnfs*6		5			n/a	5	5	5	No effect		5
	c.588dupA	p.Gln197Thrfs*7						5	n/a		No effect		5
Intron 7 (n=17)	c.588+1del	p.?		4				5	4		Total effect (Δ7)		5
	c.588+1G>T	p.?		5	5			5	5		Total effect (Δ7)	✓ [4]	5
	c.588+2T>A	p.?		4	4			5	4		Total effect (Δ7)		5
	c.588+2T>C	p.?						5		5	Total effect (Δ7)		5
	c.588+3A>G	p.?			n/a		n/a				No effect		1

	c.588+3_+6del	p.?		3	3			5	3		Total effect (Δ7)		5
	c.588+5G>A	p.?		5	5 3			5	5	5	Total effect (Δ7) [2]	✓ [5]	5
	c.588+5G>T	p.?	n/a					5	n/a		Total effect (Δ7) [3]		
	c.588+5G>C	p.?		3	5 3			5	3		Total effect (Δ7) [3]		5
	c.588+8C>A	p.?		2							No effect		1
	c.588+11G>C	p.?		1 2	1 2	n/a	n/a		1	1	No effect [2]	✓ [6]	1
	c.588+24T>C	p.?			n/a		n/a				No effect		1
	c.588+26G>A	p.?			n/a		n/a			2	No effect		1
	c.588+31G>A	p.?			n/a		n/a			3	No effect		1
	c.588+35T>A	p.?								3	No effect		1
	c.588+37T>C	p.?			n/a		n/a				No effect		1
	c.588+38G>A	p.?								3	Partial effect (Δ7)		3

Table 1. Naturally occuring *MLH1* exon 7 variants selected for this study.



C

		MES (-15%)	SSFL (-5%)	MES+ SSFL	SPiCE (11.5%)
True calls	Positives	13	12	12	13
	Negatives	5	3	4	4
	Total	**18**	**15**	**16**	**17**
False calls	Positives	0	1	0	0
	Negatives	0	1	1	0
	Total	**0**	**2**	**1**	**0**
Sensitivity		100	92	92	100
Specificity		100	75	100	50
Accuracy		100	89	94	89

Figure 1. A relatively important fraction of intronic variants alters the splicing pattern of *MLH1* exon 7 in agreement with *in silico* predictions.

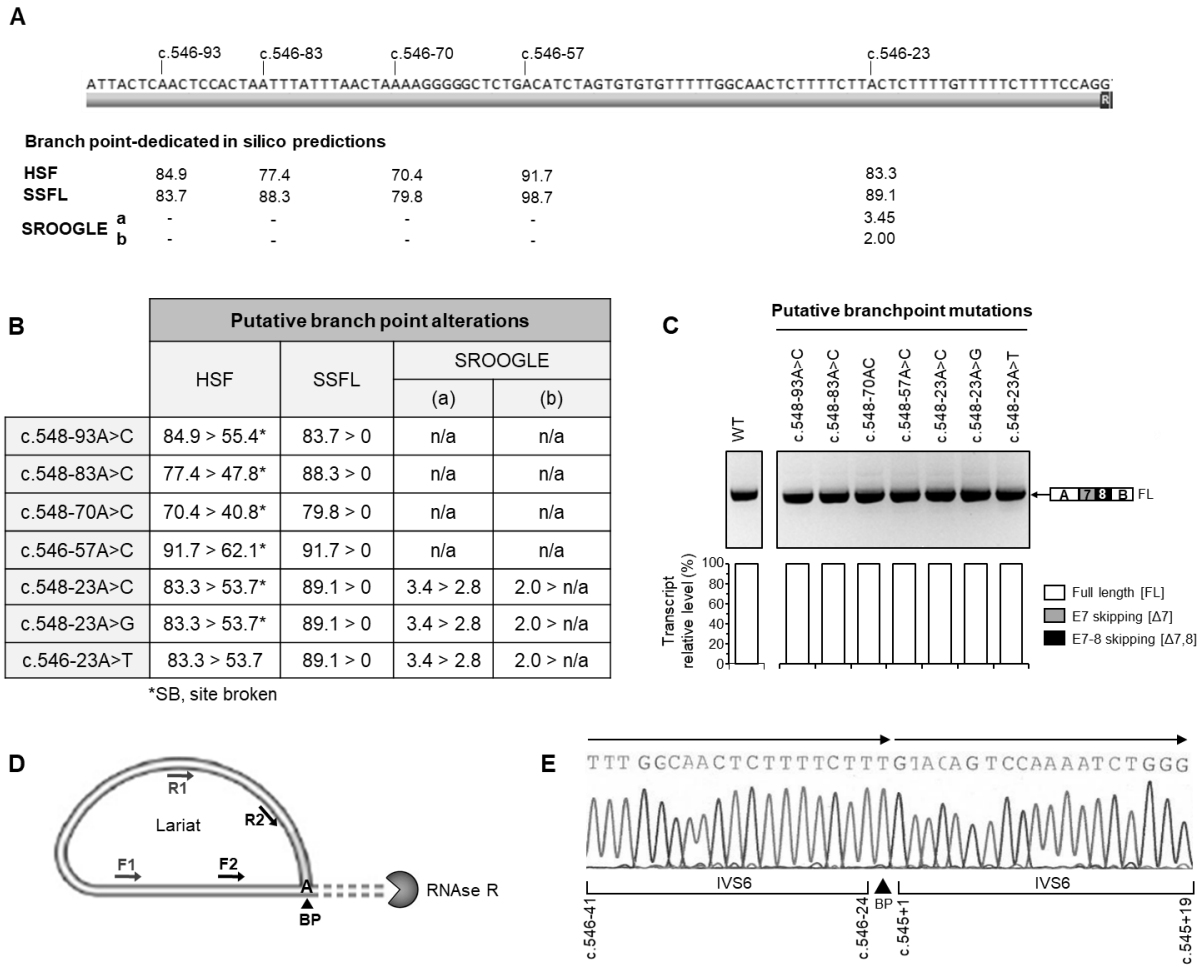
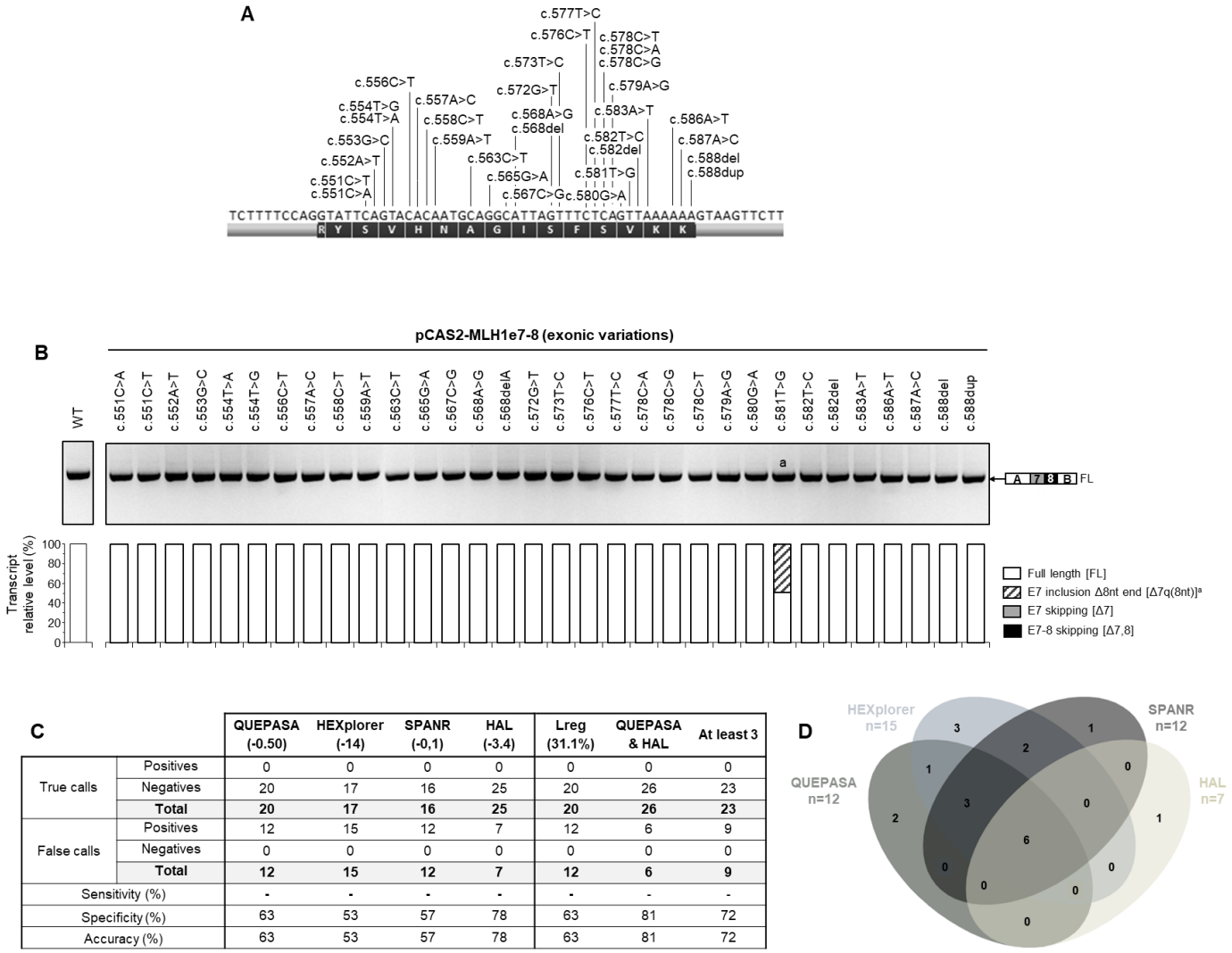


Figure 2. The adenosine at position -23 serves as branch point in *MLH1* intron 6, as predicted by bioinformatics approaches, but is dispensable for efficient splicing of exon 7.



		QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0,1)	HAL (-3.4)	Lreg (31.1%)	QUEPASA & HAL	At least 3
True calls	Positives	0	0	0	0	0	0	0
	Negatives	20	17	16	25	20	26	23
	Total	**20**	**17**	**16**	**25**	**20**	**26**	**23**
False calls	Positives	12	15	12	7	12	6	9
	Negatives	0	0	0	0	0	0	0
	Total	**12**	**15**	**12**	**7**	**12**	**6**	**9**
Sensitivity (%)		-	-	-	-	-	-	-
Specificity (%)		63	53	57	78	63	81	72
Accuracy (%)		63	53	57	78	63	81	72

Figure 3. Contrary to SRE-dedicated *in silico* predictions none of the variants mapping to *MLH1* exon 7 induce exon skipping.

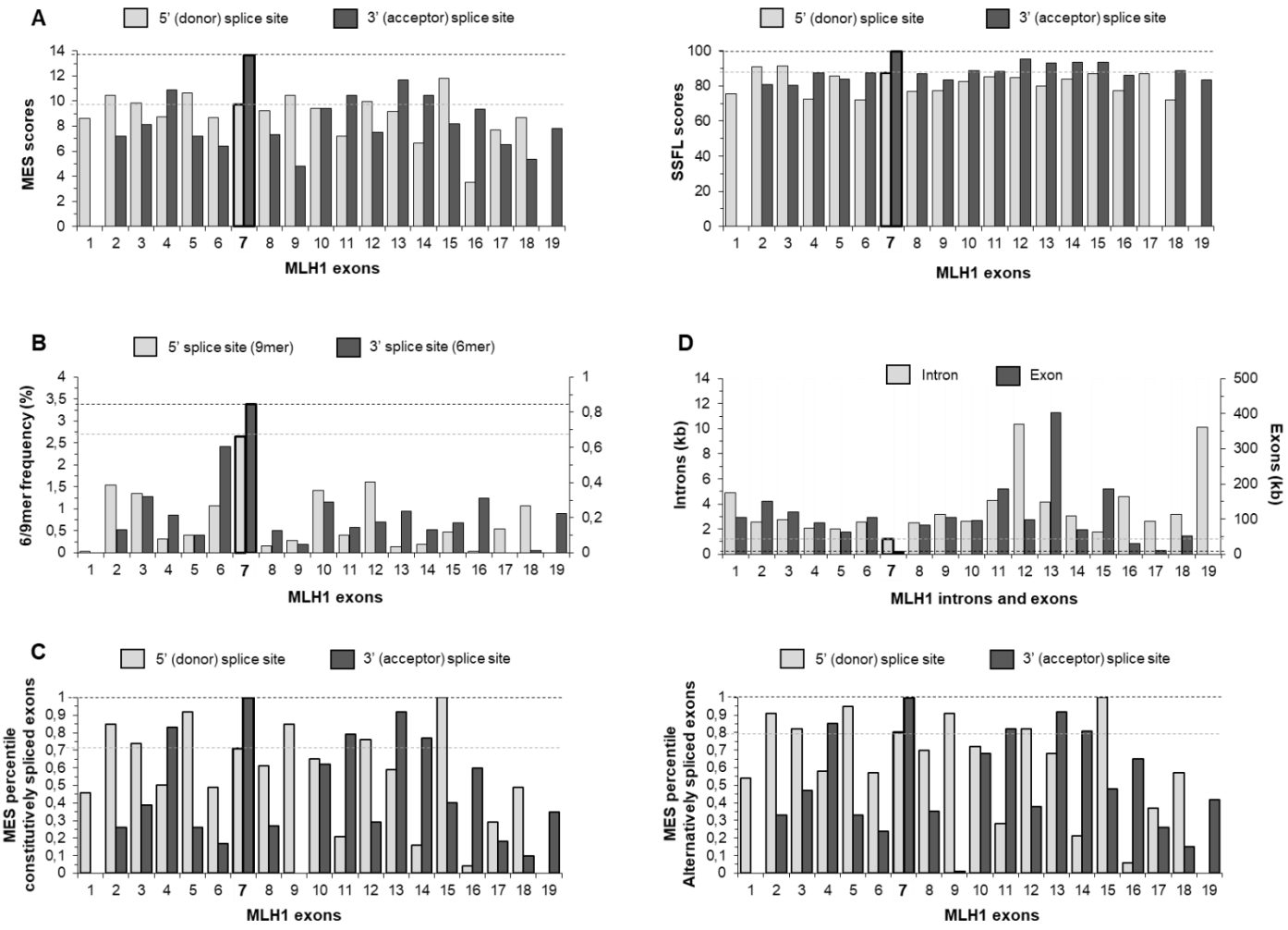
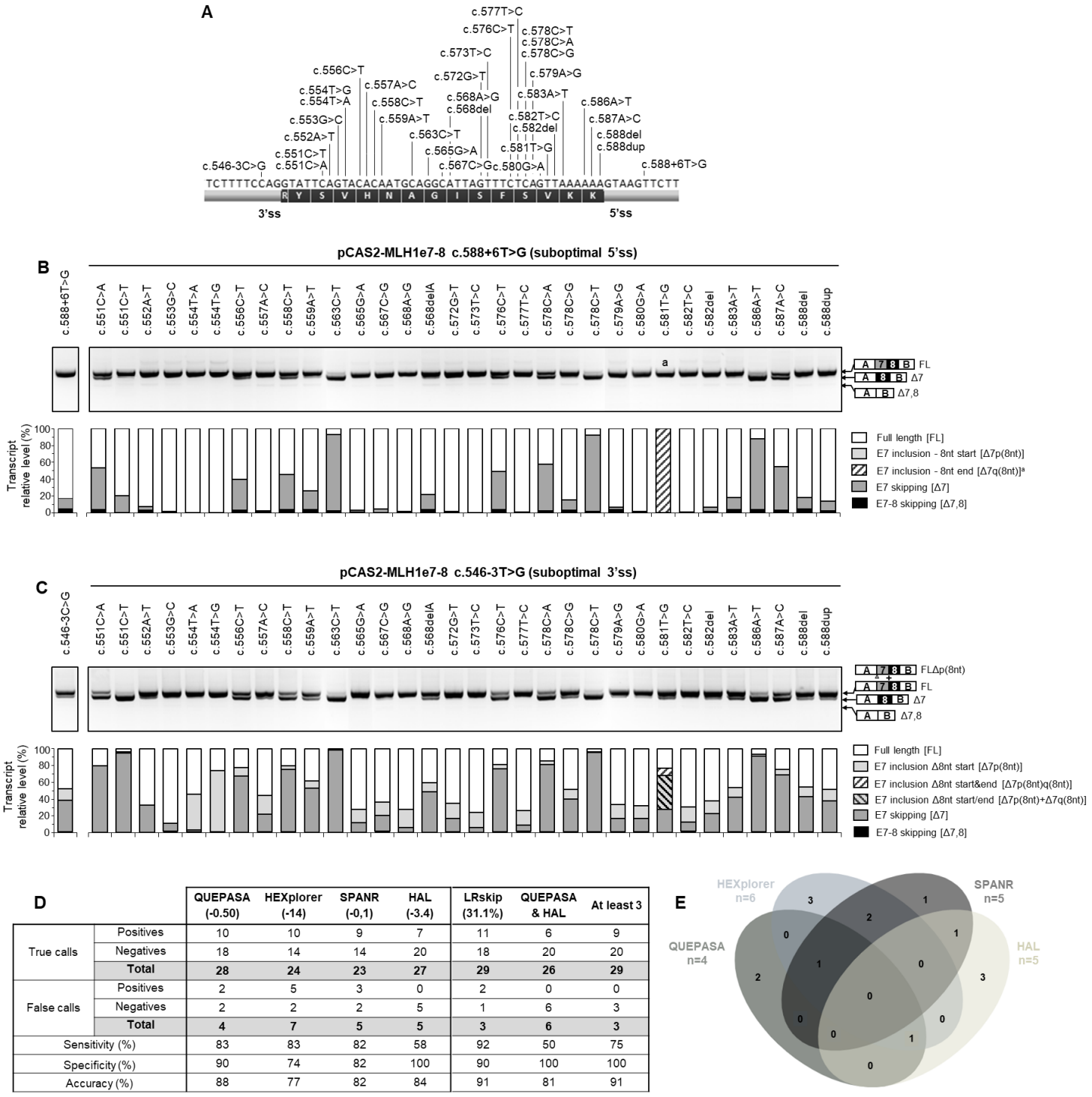


Figure 4. Features of *MLH1* exons.



D

		QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0,1)	HAL (-3.4)	LRskip (31.1%)	QUEPASA & HAL	At least 3
True calls	Positives	10	10	9	7	11	6	9
	Negatives	18	14	14	20	18	20	20
	Total	**28**	**24**	**23**	**27**	**29**	**26**	**29**
False calls	Positives	2	5	3	0	2	0	0
	Negatives	2	2	2	5	1	6	3
	Total	**4**	**7**	**5**	**5**	**3**	**6**	**3**
Sensitivity (%)		83	83	82	58	92	50	75
Specificity (%)		90	74	82	100	90	100	100
Accuracy (%)		88	77	82	84	91	81	91

Figure 5. Weakening the 5'ss or the 3'ss renders *MLH1* exon 7 sensitive to ESR mutations and unveils the predictive power of SRE-dedicated *in silico* tools.

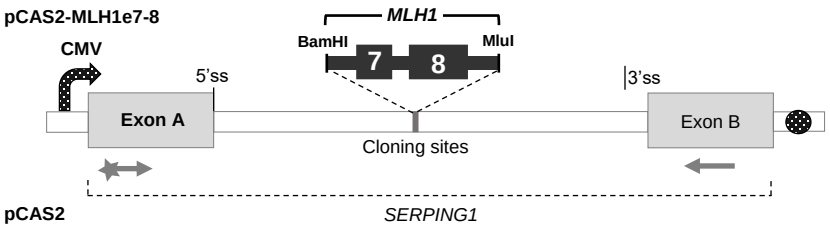
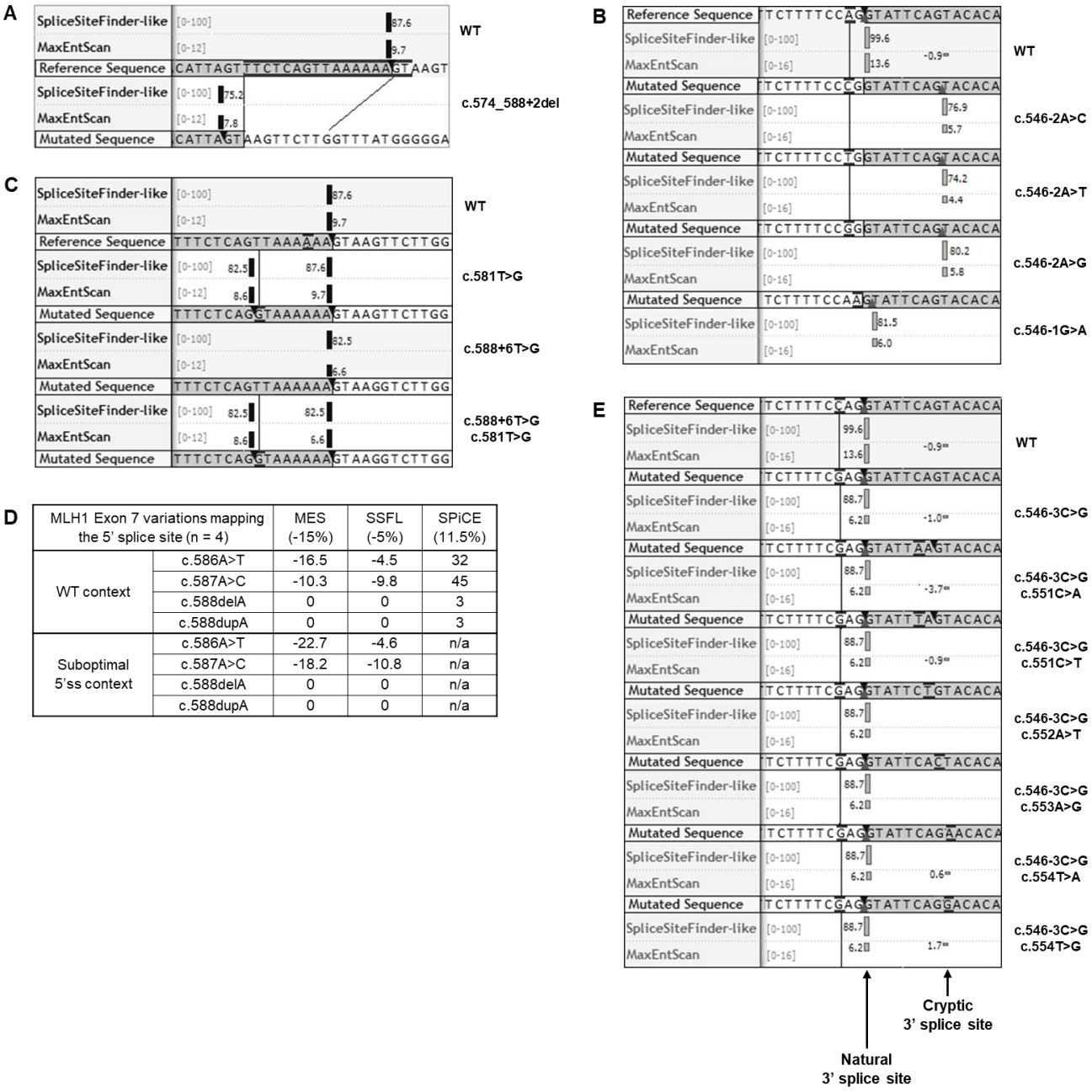


Figure S1. Structure of the pCAS2-*MLH1*e7-8 minigene used in the cell-based splicing reporter assays.



D

MLH1 Exon 7 variations mapping the 5' splice site (n = 4)		MES (-15%)	SSFL (-5%)	SPiCE (11.5%)
WT context	c.586A>T	-16.5	-4.5	32
	c.587A>C	-10.3	-9.8	45
	c.588delA	0	0	3
	c.588dupA	0	0	3
Suboptimal 5'ss context	c.586A>T	-22.7	-4.6	n/a
	c.587A>C	-18.2	-10.8	n/a
	c.588delA	0	0	n/a
	c.588dupA	0	0	n/a

Figure S2. Splice-site dedicated bioinformatics predictions of variant-induced creation of de novo splice sites or activation of cryptic splice sites experimentally detected in the minigene assays.

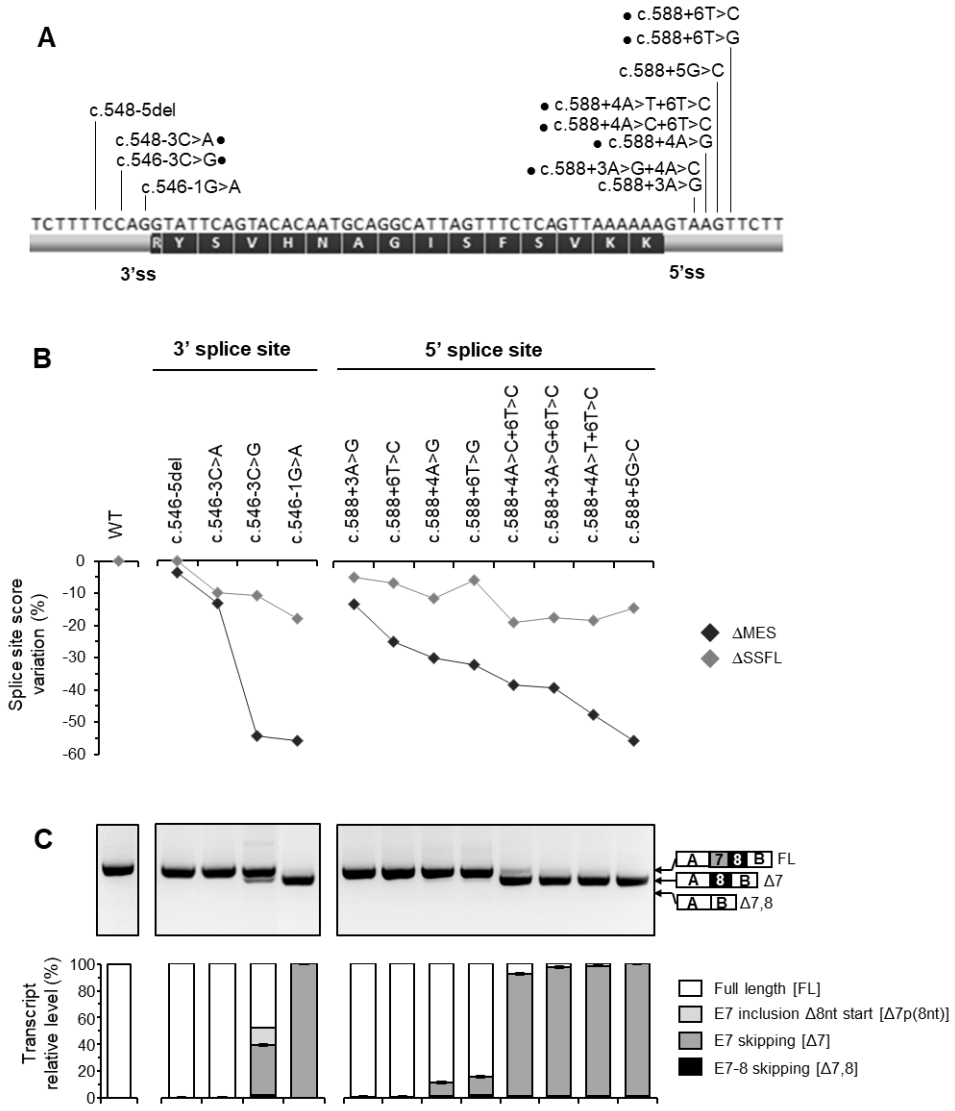


Figure S3. Impact on splicing of artificial intronic variations predicted to decrease the strength of *MLH1* exon 7 splice sites.

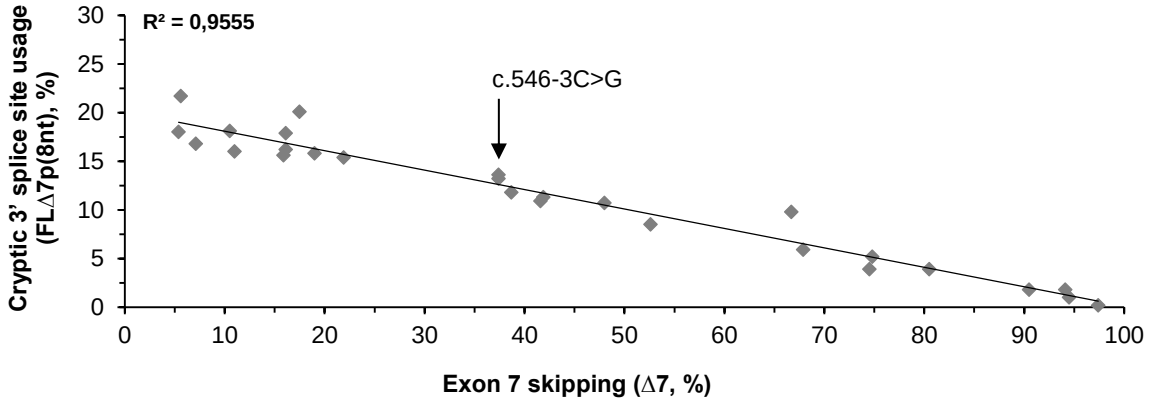


Figure S4. Correlation between variant-associated exon skipping levels and cryptic 3'ss activation obtained in the context of a suboptimal 3'ss.

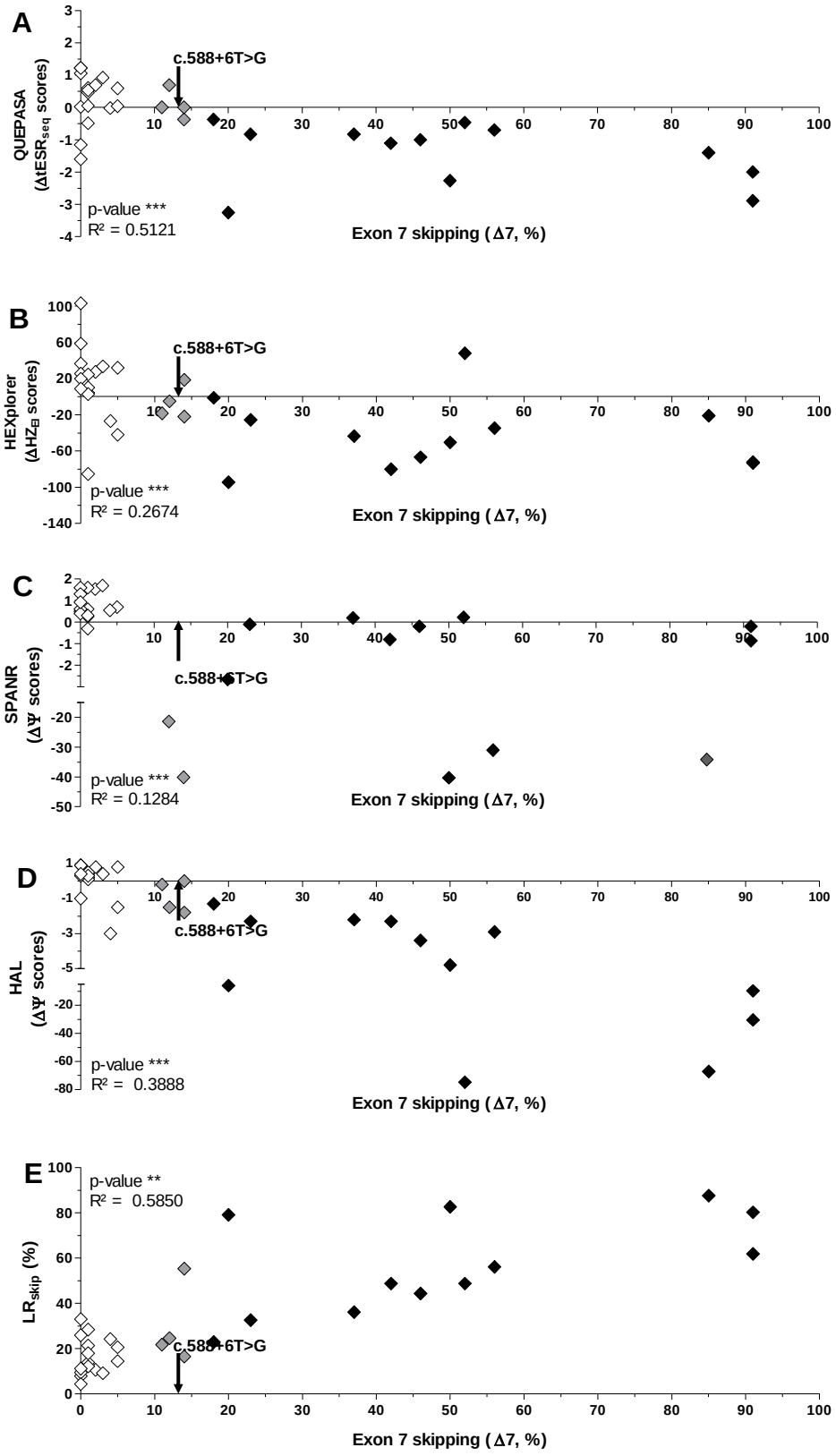


Figure S5. Correlation between variant-associated exon skipping levels described in the context of a suboptimal 5'ss and *in silico* data obtained with SRE-dedicated approaches.

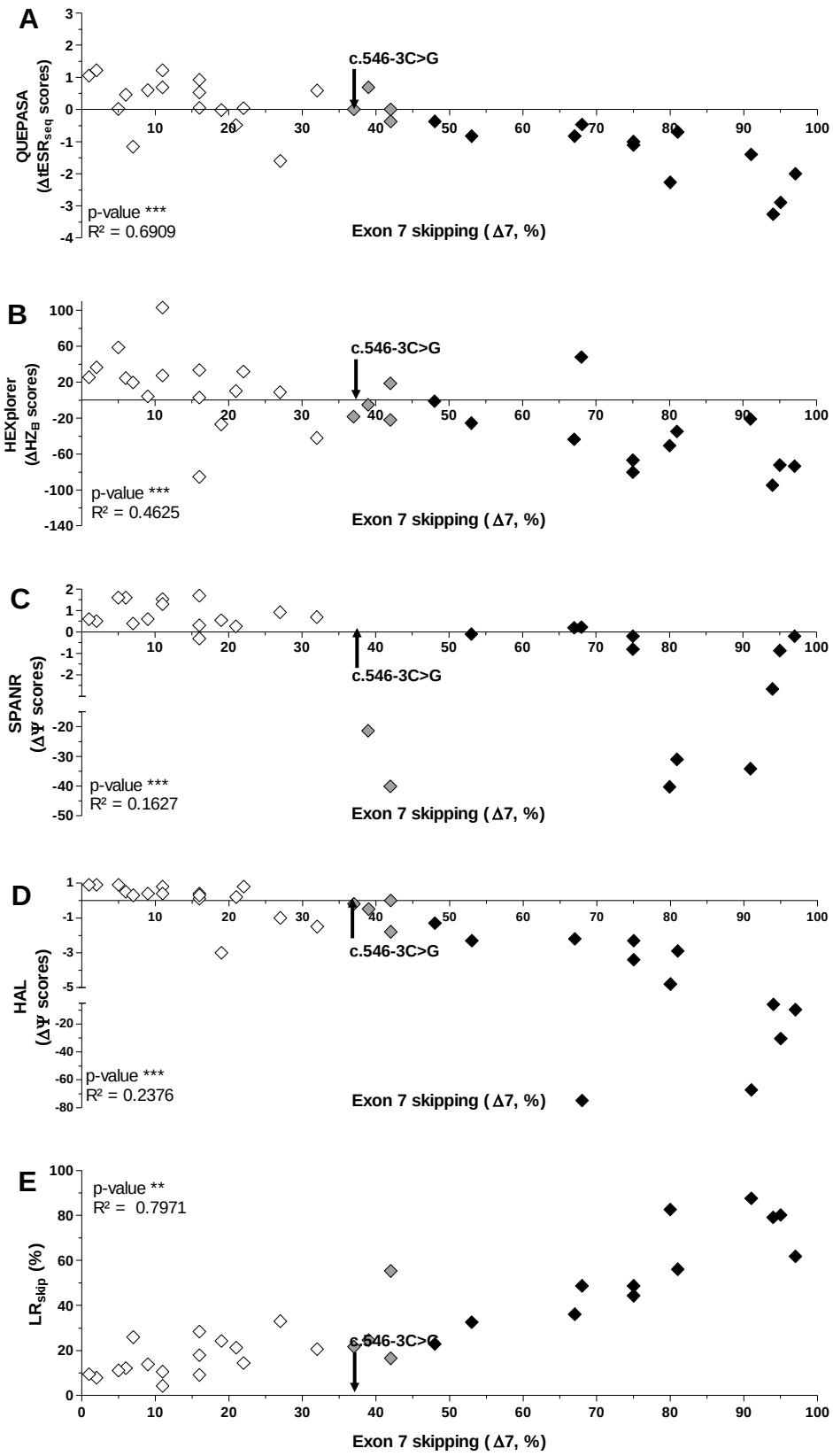


Figure S6. Correlation between variant-associated exon skipping levels described in the context of a suboptimal 3'ss and *in silico* data obtained with SRE-dedicated approaches.

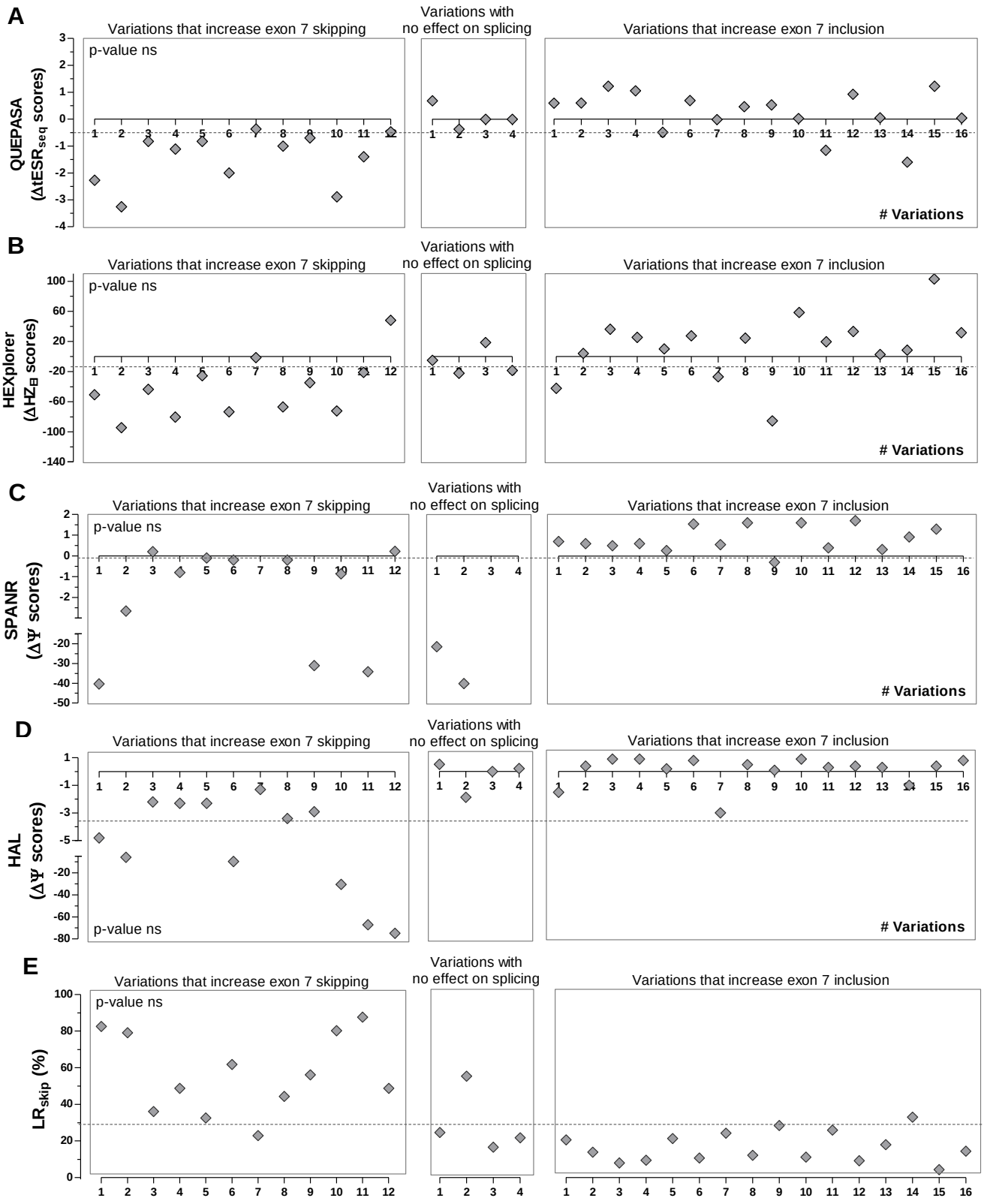


Figure S7. Comparison of the variant-associated splicing effects observed in the context of suboptimal splice sites with *in silico* data obtained with SRE-dedicated approaches.

Purpose	Template pCAS2-MLH1e7-8	Primer name[1]	Primer sequence (5'-3')[2]
PCR (cloning, minigene preparation)		MLH1Ex7-8_InFus_BamHI-F	AGGCTAAGAAGTGCAGGATCCTGGTAAAATATTAATAG GCTGTATGGAGATAT
		MLH1_Ex7-8_InFus_MluI-R	AGGGGTCAAAACAAGACGCGTGAACACATGATTCACGC CAC
Sequencing of minigene inserts		pCAS-Seq-F	GGGGTCAATAGCAGTGAGAG
		pCAS-Seq-R	GCTCCATTTCACAGGTAGAGA
RT-PCR and/or sequencing of RT-PCR products		6FAM-pCAS-KO1-F (5'-fluo)	TGACGTCGCCGCCCATCAC
		pCAS-2R	ATTGGTTGTTGAGTTGGTTGTC
Branch point mapping		M1.e7.c545-R1	CCAACTGCTCTTTCCTGGGG
		M1.e7.c545-F1	GCGTGATATCCTTGATTCTATCAGCA
		M1.e7.c546-F2	GTGGGTAAAATATTAATAGGCTGTATGGAG
		M1.e7.c546-R2	TCTGTAAGCTAGTTTTACTTTTCACCATCTA
Mutagenesis	WT	MLH1Ex7_c.546-93AC_F	GTTTAGATTACTCCACTCCAC
	WT	MLH1Ex7_c.546-83AC_F	CTCAACTCCACTACTTTATTTAAC
	WT	MLH1Ex7_c.546-70AC_F	TATTTAACTACAAGGGGGCTCTG
	WT	MLH1Ex7_c.546-57AC_F	AAGGGGGCTCTGCCATCTAGTGTG
	WT	MLH1Ex7_c.546-40TC_F	GTGTGTGTTTCTGGCAACTCTTTT
	WT	MLH1Ex7_c.546-39TG_F	GTGTGTGTTTTGGGCAACTCTTT
	WT	MLH1Ex7_c.546-35AG_F	GTGTGTTTTTGGCGACTCTTTTCTTA
	WT	MLH1Ex7_c.546-34AC_F	GTGTTTTTGGCACCTCTTTTCTTACT
	WT	MLH1Ex7_c.546-32TC_F	GTTTTTGGCAACCCTTTTCTTACTCT
	WT	MLH1Ex7_c.546-23AC_F	GCAACTCTTTTCTTCCTCTTTTGTTTTTC
	WT	MLH1Ex7_c.546-23AG-R	AAAAACAAAAGAGCAAGAAAAGAGTTGC
	WT	MLH1Ex7_c.546-23AT_F	GCAACTCTTTTCTTTCTCTTTTGTTTTTC
	WT	MLH1Ex7_c.546-18TC_F	CTCTTTTCTTACTCTCTTGTTTTTCTTTTCC
	WT	MLH1Ex7_c.546-9CG_F	CTCTTTTGTTTTTGTTTTCCAGGTATTCAG
	WT	MLH1Ex7_c.546-5del_F	CTTTTGTTTTTCTTTCCAGGTATTCAGTAC
	WT	MLH1Ex7_c.546-3CT_F	TGTTTTTCTTTTCTAGGTATTCAGTACACAA
	WT	MLH1Ex7_c.546-3CG_F	TCTTTTCGAGGTATTCAGTACACAATGC
	WT	MLH1Ex7_c.546-3CA_	TCTTTTCTAGGTATTCAGTACACAATGC
	WT	MLH1Ex7_c.546-2AC_F	TTCTTTTCCCGGTATTCAGTACACAA
	WT	MLH1Ex7_c.546-2AT_R	TGTGTACTGAATACCAGGAAAAGAAA
	WT	MLH1Ex7_c.546-1GA_F	TTTTTCTTTTCCAAGTATTCAGTACACAATG
	WT c.588+6T>G	MLH1Ex7_c.551CA_F	TTCCAGGTATTAAGTACACAATGCAGGC
	WT	MLH1Ex7_c.551CA-3CG_F	TTCGAGGTATTAAGTACACAATGCAGGC
	WT c.588+6T>G	MLH1Ex7_c.551CT_R	GCCTGCATTGTGTACTAAATACCTGGAAAAG
	WT	MLH1Ex7_c.551CT-3CG_R	GCCTGCATTGTGTACTAAATACCTCGAAAAG
	WT c.588+6T>G	MLH1Ex7_c.552AT_F	CTTTTCCAGGTATTCTGTACACAATGCAGGCA
	WT	MLH1Ex7_c.552AT-3CG_F	CTTTTCGAGGTATTCTGTACACAATGCAGGCA
	WT c.588+6T>G	MLH1Ex7_c.553GC_F	TTTTCCAGGTATTCACTACACAATGCAGGCAT
	WT	MLH1Ex7_c.553GC-3CG_F	TTTTCGAGGTATTCACTACACAATGCAGGCAT
	WT c.588+6T>G	MLH1Ex7_c.554TA_F	CCAGGTATTCAGAACACAATGCAGGCATTAG
	WT	MLH1Ex7_c.554TA-3CG_F	CGAGGTATTCAGAACACAATGCAGGCATTAG
	WT c.588+6T>G	MLH1Ex7_c.554TG_F	TTTTCCAGGTATTCAGGACACAATGCAGGCAT
	WT	MLH1Ex7_c.554TG-3CG_F	TTTTCGAGGTATTCAGGACACAATGCAGGCAT
	WT c.588+6T>G	MLH1Ex7_c.556CT_F	TCCAGGTATTCAGTATACAATGCAGGCATTAG

	WT	MLH1Ex7_c.556CT-3CG_F	TCGAGGTATTCAGTATACAATGCAGGCATTAG
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.557AC_F	GGTATTCAGTACCCAATGCAGGC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.558CT_F	GTATTCAGTACATAATGCAGGCATTAGTTTC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.559AT_F	GTATTCAGTACACTATGCAGGCATTAGTTTC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.563CT_F	ATTCAGTACACAATGTAGGCATTAGTTTCTCAGTT
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.565GA_R	GAAACTAATGCTTGCATTGTGTAC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.567CG_R	AACTGAGAAACTAATCCCTGCATTGTGTAC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.568AG_F	CAATGCAGGCGTTAGTTTCTCAGTTAAAAAAGTAAGTTC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.568delA_R	TAACTGAGAAACTAAGCCTGCATTGTGTACTG
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.572GT_R	TTTTTAACTGAGAAAATAATGCCTGCATTGTG
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.573TC_F	CAGGCATTAGCTTCTCAGTTAAAAAAGTAAGTTCTTGG
	WT	MLH1Ex7_c.573TC+6TG_F	CAGGCATTAGCTTCTCAGTTAAAAAAGTAAGGTCTTGG
	WT	c.574_588+2del_F	GCAGGCATTAGTAAGTTCTTGGTT
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.576CT_R	TAACTGAAAAACTAATGCCTGCATTGTGTACTGAATAC
	WT c.546-3C>G c.588+6T>G	MLH1Ex7_c.577TC_R	TTACTTTTTTAACTGGGAAACTAATGCCTGC
	WT c.546-3C>G	MLH1Ex7_c.578CA_R	ACTTACTTTTTTAACTTAGAAACTAATGCCTG
	WT	MLH1Ex7_c.578CA+6TG_R	CCTTACTTTTTTAACTTAGAAACTAATGCCTG
	WT c.546-3C>G	MLH1Ex7_c.578CG_R	CTTACTTTTTTAACTCAGAAACTAATGCCTGC
	WT	MLH1Ex7_c.578CG+6TG_R	CCTTACTTTTTTAACTCAGAAACTAATGCCTG
	WT c.546-3C>G	MLH1Ex7_c.578CT_R	TACTTTTTTAACTAAGAAACTAATGCCTGC
	WT	MLH1Ex7_c.578CT+6TG_R	CCTTACTTTTTTAACTAAGAAACTAATGCCTG
	WT c.546-3C>G	MLH1Ex7_c.579AG_R	AGAACTTACTTTTTTAACCGAGAAACTAATGCCTGC
	WT	MLH1Ex7_c.579AG+6TG_R	AGACCTTACTTTTTTAACCGAGAAACTAATGCCTGC
	WT c.546-3C>G	MLH1Ex7_c.580GA_R	GAACTTACTTTTTTAATTGAGAAACTAATGCC
	WT	MLH1Ex7_c.580GA+6TG_R	GACCTTACTTTTTTAATTGAGAAACTAATGCC
	WT c.546-3C>G	MLH1Ex7_c.581TG_F	GCATTAGTTTCTCAGGTAAAAAAGTAAGTTC
	WT	MLH1Ex7_c.581TG_F	GCATTAGTTTCTCAGGTAAAAAAGTAAGGTC
	WT c.546-3C>G	MLH1Ex7_c.582TC_R	AAGAACTTACTTTTTTGACTGAGAAACTAATG
	WT	MLH1Ex7_c.582TC+6TG_R	AAGACCTTACTTTTTTGACTGAGAAACTAATG

	WT c.546-3C>G	MLH1Ex7_c.582del_R	GAACTTACTTTTTTACTGAGAAACTAATGCC
	WT	MLH1Ex7_c.582del+6TG_R	GACCTTACTTTTTTACTGAGAAACTAATGCC
	WT c.546-3C>G	MLH1Ex7_c.583AT_R	CCAAGAACTTACTTTTTAAACTGAGAAACTAATGCCTG
	WT	MLH1Ex7_c.583AT+6TG_R	CCAAGACCTTACTTTTTAAACTGAGAAACTAATGCCTG
	WT c.546-3C>G	MLH1Ex7_c.586AT_R	ACCAAGAACTTACTTATTTAACTGAGAAACT
	WT	MLH1Ex7_c.586AT+6TG_R	ACCAAGACCTTACTTATTTAACTGAGAAACT
	WT c.546-3C>G	MLH1Ex7_c.587AC_R	ACCAAGAACTTACTGTTTTAACTGAGAAAC
	WT	MLH1Ex7_c.587AC+6TG_R	ACCAAGACCTTACTGTTTTAACTGAGAAAC
	WT c.546-3C>G	MLH1Ex7_c.588delA_R	TAAACCAAGAACTTACTTTTTAACTGAGAAAC
	WT	MLH1Ex7_c.588delA+6TG_R	TAAACCAAGACCTTACTTTTTAACTGAGAAAC
	WT c.546-3C>G	MLH1Ex7_c.588dupA_R	AACCAAGAACTTACTTTTTTTAACTGAGAAAC
	WT	MLH1Ex7_c.588dupA+6TG_R	AACCAAGACCTTACTTTTTTTAACTGAGAAAC
	WT	MLH1Ex7_c.588+1del_R	CCATAAACCAAGAACTTATTTTTTAACTGAGAA
	WT	MLH1Ex7_c.588+1GT_R	CATAAACCAAGAACTTTATTTTTTAACTGAGA
	WT	MLH1Ex7_c.588+2TA_R	CATAAACCAAGAACTTTCTTTTTTAACTGAGA
	WT	MLH1Ex7_c.588+2TC_R	ATAAACCAAGAACTTGCTTTTTTAACTGAG
	WT	MLH1Ex7_c.588+3AG_	ATAAACCAAGAACTCACTTTTTTAACTGAG
	WT	MLH1Ex7_c.588+3AG+4AC_F	CAGTTAAAAAAGTGCGTTCTTGGTTTATGG
	WT	MLH1Ex7_c.588+4AG_F	TCAGTTAAAAAAGTAGGTTCAAGGTTTATGG
	WT	MLH1Ex7_c.588+4AC+6TC_R	ATAAACCAAGAGCGTACTTTTTTAACTGAG
	WT	MLH1Ex7_c.588+4AT+6TC_F	CTCAGTTAAAAAAGTATGCTCTTGGTTTATG
	WT	MLH1Ex7_c.588+5GC_R	CCCATAAACCAAGAAGTTACTTTTTTAACTGAGAAACTAATG
	WT	MLH1Ex7_c.588+6TC_R	CCATAAACCAAGAGCTTACTTTTTTAACTG
	WT	MLH1Ex7_c.588+6TG_R	CCATAAACCAAGACCTTACTTTTTTAACTG
	WT	MLH1Ex7_c.588+8CA_R	CCATAAACCAATAACTTACTTTTTTAAC
	WT	MLH1Ex7_c.588+11GC_R	ATCCCCCATAAACGAAGAACTTACTT
	WT	MLH1Ex7_c.588+24TC_R	ATAAAACAAAACCGTCCCCCATAAAC
	WT	MLH1Ex7_c.588+26GA_R	TTTTCATAAAACAAAATCATCCCCCATAAAC
	WT	MLH1Ex7_c.588+31GA_R	TTTCTTTTCATAAAATAAAACCATCCCCCA
	WT	MLH1Ex7_c.588+35TA_R	CTTTTTTCTTTTCATTAAACAAAACCATCCC
	WT	MLH1Ex7_c.588+37TC_R	CCCTTTTTTCTTTTCGTAAAACAAAACCAT
	WT	MLH1Ex7_c.588+38GA_R	TCCCCTTTTTTCTTTTTATAAAACAAAACCA
PCR (cloning, minigene preparation)		MLH1Ex7-8_InFus_BamHI-F	AGGCTAAGAAGTGCAGGATCCTGGTAAAATATTAATAGGCTGTATGGAGATAT
		MLH1_Ex7-8_InFus_MluI-R	AGGGGTCAAAACAAGACGCGTGAACACATGATTCACGCCAC
Sequencing of minigene inserts		pCAS-Seq-F	GGGGTCAATAGCAGTGAGAG
		pCAS-Seq-R	GCTCCATTTCACAGGTAGAGA
RT-PCR and/or sequencing of RT-PCR products		6FAM-pCAS-KO1-F (5'-fluo)	TGACGTCGCCGCCCATCAC
		pCAS-2R	ATTGGTTGTTGAGTTGGTTGTC
Branch point mapping		M1.e7.c545-R1	CCAACTGCTCTTTCCTGGGG
		M1.e7.c545-F1	GCGTGATATCCTTGATTCTATCAGCA
		M1.e7.c546-F2	GTGGGTAAAATATTAATAGGCTGTATGGAG
		M1.e7.c546-R2	TCTGTAAGCTAGTTTTACTTTTCACCATCTA

Table S1. Description of the primers used in this study.

Purpose	Gaussian distribution (Shapiro Wilk test)	Statistical analysis	Bioinformatics approach
Linear correlation between exon inclusion levels and *in silico* predictions	Yes	Spearman	QUEPASA, HEXplorer
	No	Pearson	SPANR, HAL, LR_{skip}, LR_{inc}
Discrimination of 3 groups of variants (↑ exon skipping *versus* no effect on splicing versus ↑ exon inclusion).	Yes	Anova (Bonferroni post-tests)	QUEPASA, HEXplorer
	No	Kruskal-Wallis (Duns post-tests)	SPANR, HAL, LR_{skip}, LR_{inc}
Discrimination of 2 groups of variants (variants that increase exon skipping or exon inclusion versus those that do not)	Yes	Student	QUEPASA, HEXplorer
	Yes	Student with Welsh's correction	
	No	Mann-Whitney	SPANR, HAL, LR_{skip}, LR_{inc}

Table S2. Description of statistical analyses conducted in this study.

MLH1 intronic variants (n = 31)	Effect on splicing	Exon 7 inclusion (%)	ΔMES (-15%)	ΔSSFL (-5%)	MES+ SSFL	SPiCE (11.5%)
WT	-	0	-	-		-
c.546-2A>C	↑ Skipping	4	-100	-100	+	100
c.546-2A>G	↑ Skipping	0	-100	-100	+	100
c.546-2A>T	↑ Skipping	0	-100	-100	+	100
c.546-1G>A	↑ Skipping	0	-100	-100	+	100
c.574_588+2del	*De novo* 5'ss	0	-100	-100	+	100
c.588+1del	↑ Skipping	0	-100	-100	+	100
c.588+1G>T	↑ Skipping	0	-100	-100	+	100
c.588+2T>A	↑ Skipping	0	-100	-100	+	100
c.588+2T>C	↑ Skipping	1	-100	-1	-	99
c.588+3_+6del	↑ Skipping	0	-100	-100	+	1
c.588+5G>A	↑ Skipping	3	-58	-14	+	99.7
c.588+5G>T	↑ Skipping	1	-62	-14	+	99.8
c.588+5G>C	↑ Skipping	0	-55	-15	+	99.7
c.588+38G>A	↑ Skipping	92	n/a	n/a	n/a	n/a
c.546-40T>C	No effect	100	n/a	n/a	n/a	n/a
c.546-39T>G	No effect	100	n/a	n/a	n/a	n/a
c.546-35A>G	No effect	100	n/a	n/a	n/a	n/a
c.546-34A>C	No effect	100	n/a	n/a	n/a	n/a
c.546-32T>C	No effect	100	n/a	n/a	n/a	n/a
c.546-18T>C	No effect	100	-2	n/a	n/a	n/a
c.546-9C>G	No effect	100	-9	0	-	5.5
c.546-5del	No effect	100	-4	0	-	3.6
c.546-3C>T	No effect	100	+1	-6	-	21.8
c.588+3A>G	No effect	100	-14	-4	-	19.1
c.588+8C>A	No effect	100	n/a	n/a	n/a	n/a
c.588+11G>C	No effect	100	n/a	n/a	n/a	n/a
c.588+24T>C	No effect	100	n/a	n/a	n/a	n/a
c.588+26G>A	No effect	100	n/a	n/a	n/a	n/a
c.588+31G>A	No effect	100	n/a	n/a	n/a	n/a
c.588+35T>A	No effect	100	n/a	n/a	n/a	n/a
c.588+37T>C	No effect	100	n/a	n/a	n/a	n/a
True calls	Positives		13	12	12	13
	Negatives		5	3	4	4
	Total		**18**	**15**	**16**	**17**
False calls	Positives		0	1	0	0
	Negatives		0	1	1	0
	Total		**0**	**2**	**1**	**0**
Sensitivity			100	92	92	100
Specificity			100	75	100	100
Accuracy			100	88	94	100

Table S3. Comparison of intronic variant-associated splicing effects observed in the pCAS2-*MLH1*e7-8 minigenes and associated splice site-dedicated bioinformatics approaches.

Variations *MLH1* Exon 7 (n = 32)	Effect on splicing	Exon 7 skipping (%)	QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0.1)	HAL (-3.4)	LR_{skip} (31.1%)	QUEPASA & HAL	At least 3
WT	-	0	-	-	-	-	-	-	-
c.551C>A	No effect	0	-2.27	-50.7	-40.3	-4.8	82.6	2/2	4/4
c.551C>T	No effect	0	-3.26	-94.6	-2.66	-5.9	79.1	2/2	4/4
c.552A>T	No effect	0	0.59	-42.2	0.7	-1.5	20.6	0/2	1/4
c.553G>C	No effect	0	0.60	4.1	0.6	0.4	13.8	0/2	0/4
c.554T>A	No effect	0	1.22	36.3	0.5	0.9	7.9	0/2	0/4
c.554T>G	No effect	0	1.05	25.4	0.6	0.9	9.4	0/2	0/4
c.556C>T	No effect	0	-0.83	-43.7	0.2	-2.2	36.1	1/2	2/4
c.557A>C	No effect	0	-0.49	10.1	0.26	0.2	21.3	0/2	0/4
c.558C>T	No effect	0	-1.11	-80.3	-0.8	-2.3	48.7	1/2	3/4
c.559A>T	No effect	0	-0.83	-25.6	-0.1	-2.3	32.5	1/2	3/4
c.563C>T	No effect	0	-2.00	-73.5	-0.2	-9.6	61.8	0/2	4/4
c.565G>A	No effect	0	0.69	27.4	1.54	0.8	10.6	0/2	0/4
c.567C>G	No effect	0	-0.02	-27.0	0.55	-3	24.2	0/2	1/4
c.568A>G	No effect	0	0.46	24.5	1.6	0.5	12.1	0/2	0/4
c.568delA	No effect	0	-0.37	-1.4	n/a	-1.3	22.9	0/2	0/4
c.572G>T	No effect	0	0.53	-85.5	-0.3	0.1	28.4	0/2	2/4
c.573T>C	No effect	0	0.02	58.5	1.6	0.9	11.1	0/2	0/4
c.576C>T	No effect	0	-1.00	-66.9	-0.2	-3.4	44.3	2/2	4/4
c.577T>C	No effect	0	-1.16	19.6	0.4	0.3	25.9	1/2	1/4
c.578C>A	No effect	0	-0.70	-34.8	-31	-2.9	56.1	1/2	3/4
c.578C>T	No effect	0	-2.89	-72.3	-0.87	-30.4	80.2	2/2	4/4
c.578C>G	No effect	0	0.69	-5.2	-21.4	0.5	24.6	0/2	1/4
c.579A>G	No effect	0	0.92	33.3	1.7	0.4	9.1	0/2	0/4
c.580G>A	No effect	0	0.05	2.7	0.31	0.3	17.9	0/2	0/4
c.582T>C	No effect	0	1.22	102.9	1.3	0.4	4.3	0/2	0/4
c.581T>G*	FLΔq(8nt)	0	-1.60	8.6	0.92	-1.0	33.0	1/2	1/4
c.582del	No effect	0	0.04	31.7	n/a	0.8	14.4	0/2	0/4
c.583A>T	No effect	0	-0.37	-22.2	-40.1	-1.8	55.3	0/2	2/4
c.586A>T*	No effect	0	-1.40	-21.1	-34.16	-67.1	87.6	2/2	4/4
c.587A>C*	No effect	0	-0.47	48.0	0.22	-74.9	48.7	1/2	1/4
c.588delA*	No effect	0	0.00	18.5	n/a	0	16.5	0/2	0/4
c.588dupA*	No effect	0	0.00	-18.5	n/a	0.2	21.7	0/2	1/4
True calls	Positives		0	0	0	0	0	0	0
	Negatives		20	17	16	25	20	26	23
	Total		**20**	**17**	**16**	**25**	**20**	**26**	**23**
False calls	Positives		12	15	12	7	12	6	9
	Negatives		0	0	0	0	0	0	0
	Total		**12**	**15**	**12**	**7**	**12**	**6**	**9**
Sensitivity			-	-	-	-	-	-	-
Specificity			63	53	57	78	63	81	72
Accuracy			63	53	57	78	63	81	72

Table S4. Comparison of exonic variant-associated splicing effects observed in the pCAS2-*MLH1*e7-8 minigenes and associated SRE-dedicated *in silico* predictions.

Variations *MLH1* Exon 7 (n = 32)	Effect on splicing	Exon 7 skipping (%)		QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0.1)	HAL (-3.4)	LR_{skip} (31.1%)	QUEPASA & HAL	At least 3
c.546-3C>G	-	37	-	-	-	-	-	-	-	-
c.588+6T>G	-		13	-	-	-	-	-	-	-
c.551C>A	↑ Skipping	80	50	-2.27	-50.7	-40.3	-4.8	82.6	2/2	4/4
c.551C>T	↑ Skipping	94	20	-3.26	-94.6	-2.66	-5.9	79.1	2/2	4/4
c.556C>T	↑ Skipping	67	37	-0.83	-43.7	0.2	-2.2	36.1	1/2	2/4
c.558C>T	↑ Skipping	75	42	-1.11	-80.3	-0.8	-2.3	48.7	1/2	3/4
c.559A>T	↑ Skipping	53	23	-0.83	-25.6	-0.1	-2.3	32.5	1/2	3/4
c.563C>T	↑ Skipping	97	91	-2.00	-73.5	-0.2	-9.6	61.8	2/2	4/4
c.568delA	↑ Skipping	48	18	-0.37	-1.4	n/a	-1.3	22.9	0/2	0/4
c.576C>T	↑ Skipping	75	46	-1.00	-66.9	-0.2	-3.4	44.3	2/2	4/4
c.578C>A	↑ Skipping	81	56	-0.70	-34.8	-31	-2.9	56.1	1/2	3/4
c.578C>T	↑ Skipping	95	91	-2.89	-72.3	-0.87	-30.4	80.2	2/2	4/4
c.586A>T*	↑ Skipping	91	85	-1.40	-21.1	-34.16	-67.1	87.6	2/2	4/4
c.587A>C*	↑ Skipping	68	52	-0.47	48.0	0.22	-74.9	48.7	1/2	1/4
c.578C>G	No effect	39	12	0.69	-5.2	-21.4	0.5	24.6	0/2	1/4
c.583A>T	No effect	42	14	-0.37	-22.2	-40.1	-1.8	55.3	0/2	2/4
c.588delA*	No effect	42	14	0.00	18.5	n/a	0	16.5	0/2	0/4
c.588dupA*	No effect	37	11	0.00	-18.5	n/a	0.2	21.7	0/2	1/4
c.552A>T	↑ Inclusion	32	5	0.59	-42.2	0.7	-1.5	20.6	0/2	1/4
c.553G>C	↑ Inclusion	9	1	0.60	4.1	0.6	0.4	13.8	0/2	0/4
c.554T>A	↑ Inclusion	2	0	1.22	36.3	0.5	0.9	7.9	0/2	0/4
c.554T>G	↑ Inclusion	1	0	1.05	25.4	0.6	0.9	9.4	0/2	0/4
c.557A>C	↑ Inclusion	21	1	-0.49	10.1	0.26	0.2	21.3	0/2	0/4
c.565G>A	↑ Inclusion	11	2	0.69	27.4	1.54	0.8	10.6	0/2	0/4
c.567C>G	↑ Inclusion	19	4	-0.02	-27.0	0.55	-3	24.2	0/2	1/4
c.568A>G	↑ Inclusion	6	1	0.46	24.5	1.6	0.5	12.1	0/2	0/4
c.572G>T	↑ Inclusion	16	1	0.53	-85.5	-0.3	0.1	28.4	0/2	2/4
c.573T>C	↑ Inclusion	5	0	0.02	58.5	1.6	0.9	11.1	0/2	0/4
c.577T>C	↑ Inclusion	7	0	-1.16	19.6	0.4	0.3	25.9	1/2	1/4
c.579A>G	↑ Inclusion	16	3	0.92	33.3	1.7	0.4	9.1	0/2	0/4
c.580G>A	↑ Inclusion	16	1	0.05	2.7	0.31	0.3	17.9	0/2	0/4
c.581T>G*	FLΔq(8nt)	27	0	-1.60	8.6	0.92	-1.0	33.0	1/2	1/4
c.582T>C	↑ Inclusion	11	0	1.22	102.9	1.3	0.4	4.3	0/2	0/4
c.582del	↑ Inclusion	22	5	0.04	31.7	n/a	0.8	14.4	0/2	0/4
True calls	Positives			10	10	9	7	11	6	9
	Negatives			18	14	14	20	18	20	20
	Total			**28**	**24**	**23**	**27**	**29**	**26**	**29**
False calls	Positives			2	5	3	0	2	0	0
	Negatives			2	2	2	5	1	6	3
	Total			**4**	**7**	**5**	**5**	**3**	**6**	**3**
Specificity				83	83	82	58	92	50	75
Sensitivity				90	74	82	100	90	100	100
Accuracy				88	77	82	84	91	81	91

Table S5. Comparison of exonic variant-associated splicing effects observed in the context of suboptimal splice sites and associated SRE-dedicated *in silico* predictions.

CHAPITRE III : PROBLEMATIQUE DE L'INTERPRETATION DES MUTATIONS D'EPISSAGE A EFFET PARTIEL

L'identification de la mutation causale est essentielle au diagnostic moléculaire et à l'optimisation de la prise en charge des patients et de leurs apparentés atteint d'un syndrome seins-ovaires. A ce jour, même si la plupart des altérations constitutionnelles à l'origine d'un syndrome seins-ovaires ont été identifiées dans les gènes *BRCA,* de nombreux cas évocateurs restent inexpliqués, soit parce qu'aucune mutation n'est observée dans ces gènes, soit parce que des VSI sont détectés. Les VSI, qui peuvent représenter jusqu'à 60% des variations détectées dans les gènes BRCA, représentent un obstacle majeur au diagnostic moléculaire du syndrome seins-ovaires, et particulièrement en cette période de médecine personnalisée, où l'identification de la mutation pathogène conditionne l'accès à certaines thérapies ciblées, notamment celles basées sur l'utilisation des PARPi, récemment mises en place dans le traitement des tumeurs déficientes en BRCA.

Les travaux de menés au sein de notre unité ont permis de mettre en évidence à l'aide des tests fonctionnels d'épissage basés sur l'utilisation de minigènes et l'étude de l'ARN des patients, qu'une fraction importante des VSI détectées dans les gènes BRCA, provoque des défauts d'épissage, hors phase pour la plupart, permettant de classer ces VSI comme pathogènes. Cependant, certaines VSI entrainent des modifications en phase, celles localisées dans l'exon 3 de *BRCA2,* contenant le domaine d'interaction avec la protéine PALB2, impliqué dans la recombinaison homologue et donc essentiel à l'activité de la protéine BRCA2. Si certaines variations à l'origine d'un Δ3 total ont pu être classées comme délétères, et d'autres, augmentant très partiellement le Δ3 ont pu être classées comme neutres sur la base de données génétiques et/ou fonctionnelles, la plupart des variations responsables d'un Δ3 partiel restent actuellement considérées comme des VSI et ne peuvent être utilisées au titre du diagnostic. Nous avons fait l'hypothèse qu'il pourrait exister un seuil de pathogénicité à partir duquel une mutation entrainant une production trop importante de Δ3 serait délétère.

Pour tester cette hypothèse, nous avons utilisé comme modèle d'étude des variations à l'origine d'un saut de l'exon 3 d'intensité croissante (15 à 75%), mais silencieuses sur le plan traductionnel (variations introniques ou synonymes), que nous avons identifiées à l'aide d'une approche combinant les prédictions bioinformatiques dédiées à l'épissage et des tests fonctionnels d'épissage basé sur l'utilisation de minigènes et l'analyse du matériel biologique du patient, lorsque disponible. Notre stratégie a consisté à évaluer, grâce à un test de complémentation basé sur l'utilisation de cellules souches embryonnaires de souris visant à évaluer la capacité d'une variation à restaurer l'activité de BRCA2, l'effet combiné de ces variations sur l'ARN et sur la protéine, dans un contexte génomique, biologique et cellulaire proche du contexte naturel. Afin de valider notre système, nous avons également inclus deux variations contrôles, responsables d'un Δ3 très faible (7%) ou total (95%) et actuellement classées comme neutre et délétère, respectivement, sur la base de données cliniques, génétiques et familiales.

Nos résultats basés sur l'utilisation des cellules souches ont permis de confirmer, dans un premier temps, le caractère neutre et délétère des mutations contrôles initialement classées à l'aide de données génétiques, cliniques et familiales, et de reproduire les défauts d'épissage observés dans le « test minigène », validant ainsi notre système. Nos données indiquent également que l'ensemble des cellules porteuses des VSI responsables d'un Δ3 partiel dans le test minigène à hauteur de 15-65%, montrent une viabilité similaire à celle des cellules WT et des cellules porteuses de la variation neutre, indiquant que ces mutations ne sont pas délétères. En effet, des tests fonctionnels basés sur l'étude de la sensibilité des cellules aux agents génotoxiques et aux irradiations ont permis de confirmer la caractère neutre de ces variations. En revanche, les variations induisant un Δ3 partiel de l'ordre de 75% sont associées à une réduction sévère de la viabilité et à une sensibilité accrue aux agents génotoxiques et aux irradiations, suggérant que ces mutations altèrent la fonction de BRCA2. Des analyses complémentaires, basées notamment sur la collecte des données génétiques, cliniques, tumorales et familiales des patientes porteuses des variations naturelles étudiées sont en cours afin de réévaluer le caractère pathogène de ces variations et de statuer sur leur pathogénicité.

L'ensemble de ces données soulèvent l'existence d'un seuil de pathogénicité à partir duquel une mutation entrainant une réduction trop importante de transcrits pleine longueur serait délétère. En d'autre terme, nos résultats suggèrent qu'un allèle BRCA2 produisant, en minigène, jusqu'à

environ 75% de transcrits codant une protéine BRCA2 déficiente ne confère pas nécessairement un risque élevé de cancer, indiquant que l'activité tumeur-suppressive BRCA2 tolère une réduction substantielle du niveau d'expression. L'étude de ce modèle contribue à l'interprétation clinique de VSI conduisant à différents niveaux de Δ3 et plus généralement de VSI conduisant à des défauts d'épissage partiels.

Ces travaux portant sur l'exon 3 de BRCA2 sont actuellement en cours.

Calibration of pathogenicity of partial splicing defects: The model of *BRCA2* Exon 3

Tubeuf H[1,2], Castelain G[1], Sullivan T[3], Caputo S[4], Southon E[3], North SL[3], Cleveland L[3], Abdat J[1], Rondeau J[1], Tourneur S[1], Frebourg T[5], Gaildrat P[1], Sharan SK[3], Martins A[1]

1. Inserm-U1245, UNIROUEN, Normandie University, Normandy Centre for Genomic and Personalized Medicine, Rouen, France. **2** Interactive Biofotware, Rouen, France. **3**. Mouse Cancer Genetics Program, Center for Cancer Research, National Cancer Institute, Frederick, MD, 21702, USA. **4** Institut Curie, Service de Génétique, Paris, France. **5** Department of Genetics, University Hospital, Normandy Centre for Genomic and Personalized Medicine, Rouen, France.

Author contributions

AM and HT conceived and designed the project. HT generated the experimental, bioinformatics and statistical data. SC performed familial data analysis. HT, SKS and AM were involved in data interpretation. HT, SKS and AM wrote the manuscript. All authors read and approved the final manuscript.

Fundings

This project was supported by the OpenHealth Institute, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (Gefluc) as well as the European Union and Région Normandie. Europe gets involved in Normandie with European Regional Development Fund (ERDF). HT was funded by a CIFRE PhD fellowship (#2015/0335) from the French Association Nationale de la Recherche et de la Technologie in the context of public-private partnership between INSERM and Interactive Biosoftware.

Introduction

Approximately 10% of women diagnosed with breast and/or ovarian cancer report a strong family history, motivating the genetic screening for breast and ovarian cancer susceptibility genes. Hereditary breast and ovarian cancers (HBOC) are mainly due to germline pathogenic mutation within *BRCA1* (MIM #113705) and *BRCA2* (MIM #600185), which are responsible for about 25% of HBOC cases (Kast *et al.*, 2016). Since the identification of inherited *BRCA1* or *BRCA2* mutations as a major increased lifetime risk of developing breast (57–65% and 45-55%) and/or ovarian cancer (39–44% and 11-18% risk by age 70 years) (Nielsen *et al.*, 2016), genetic testing for variants in *BRCA1/2* got increasingly important. Indeed, the identification of a pathogenic mutation is essential for molecular diagnosis and cancer-risk-management strategies, including intensified screening programs, presymptomatic genetic testing of family members, enhanced cancer surveillance, the option to undergo preventive prophylactic surgery and/or to benefit from targeted therapies such as PARP inhibitors recommended in the treatment of BRCA-deficient tumors (Ledermann *et al.*, 2014; Lindor *et al.*, 2013). However, the functional impact of many variants identified within *BRCA1/2* by genetic testing remains unknown. Those variants, so called variations of unknown significance (VUS), represent up to 50% of the detected variations in the BRCA genes, and are considered as one of the major obstacles to the molecular diagnosis of HBOC and to optimal patient care (Caputo *et al.*, 2012; Eccles *et al.*, 2015; Landrum *et al.*, 2014; Szabo *et al.*, 2000).

VUS are generally typified by missense SNVs, but also correspond to silent substitutions, small in-frame insertions and deletions as well as intronic variants, all of which can potentially affect pre-mRNA splicing (Spurdle *et al.*, 2012). Mutations that affect mRNA splicing account for at least 15% of all disease-causing point mutations with up to 50% of all mutations detected in some genes (Ars *et al.*, 2000; Baralle *et al.*, 2009; Teraoka *et al.*, 1999). Numerous studies indicate that a large fraction of VUS, including those detected in the *BRCA* genes, induce RNA splicing defects as determined by performing patient RNA analyses and/or cell-based minigene splicing assays (Acedo *et al.*, 2012; Bonnet *et al.*, 2008; Di Giacomo *et al.*, 2013; Gaildrat *et al.*, 2012; Houdayer *et al.*, 2012; Leman *et al.*, 2018; Sanz *et al.*, 2010; Théry *et al.*, 2011). Many of these splicing defects results in loss of function by introduction of a premature termination codon, allowing to classify these variations as pathogenic (Walker *et al.*, 2013). However, some variations lead to

either in-frame modifications and/or partial (leaky) splicing defects. If accumulating evidence has suggested that some of them, which are permissive to embryonic lethality normally induced by BRCA1 or BRCA2 biallelic loss, may be hypomorphic (Biswas *et al.*, 2011; Meulemans *et al.*, in preparation), the consequences of most of these splicing anomalies on the resulting BRCA1 and BRCA2 protein function remain unknown. Thus, such variations remain classified as VUS and cannot be used for clinical decisions (Walker *et al.*, 2013). This is such the case of several mutations occurring in the exon 3 of *BRCA2*.

The third exon of *BRCA2* has been shown to comprise (i) the transactivation core bound in particular by EMSY, a negative regulator of BRCA2 transcription activation potential and (ii) the PALB2/FANCN-interaction domain which links BRCA1 to BRCA2 though PALB2 (Martinez *et al.*, 2015). This interaction is critical for the recruitment of BRCA2 to double stand breaks (DSBs) and for its role in homologous recombination (HR) (Prakash *et al.*, 2015). Indeed, missense variants that disrupt this interaction, such as W31R and W31C, have been functionally shown to alter BRCA2 function by reducing drastically HR-mediated DSB repair rendering cells sensitive to DNA damage and have been reported as deleterious (Biswas *et al.*, 2012). In addition, deletion of exon 3 as well as several variations inducing total exon 3 skipping are associated with increased risk of breast and ovarian cancer and have been classified as pathogenic on the basis of co-segregation analysis and functional data (Caputo *et al.*, 2018; Muller *et al.*, 2011). Altogether, these data demonstrate the functional importance of BRCA2 exon 3. However, there is a low proportion of BRCA2 physiological alternative transcripts lacking the exon 3 (Δ3) that has been detected in normal tissues, including mammary glands and prostate tissues (Muller *et al.*, 2011; Zou *et al.*, 1999). Furthermore, it has been recently demonstrated the nonpathogenicity of the BRCA2 c.68-7T>A variation which is not associated with increased risk of breast cancer (OR 1.03) despite a low increased in Δ3 transcript level (Colombo *et al.*, 2018). Taken together, these data raise the hypothesis of a threshold for pathogenicity of Δ3 transcripts level above which the function of BRCA2 is impaired leading to an increased predisposition to breast and ovarian cancers.

In order to better understand the contribution to pathogenicity of variant-induced partial *BRCA2* exon 3 skipping, we identified translationally silent variations (synonymous and intronic) causing increasing levels of exon 3 skipping ("gradient") by resorting both to splicing-dedicated *in silico* analyses and functional splicing assay based on either minigene or patient's RNA when available.

We then intended to assess the functional consequences of selected BRCA2 variants by combining multifactorial likelihood analysis, including genetic, clinical, tumoral, cosegregation and familial data, with functional data from a mouse embryonic stem cell (mESC)-based complementation assay used to evaluate the combined effect of variations on RNA splicing and protein function. Our study showed that BRCA2 allele producing as much as 70% of Δ3 transcript encoding deficient protein may not necessarily confer high-risk of developing cancer or may be associated with lower penetrance and provide evidence that BRCA2 tumor suppressor activity tolerates a substantial reduction in expression level. Altogether, these data contribute to the clinical interpretation of a number of VUS within BRCA2 leading to partial exon 3 skipping and to the re-evaluation of the pathogenicity of certain mutations currently considered deleterious. Characterization of these mutations improves the molecular diagnosis and medical management of patients and their relatives, with a direct benefit for HBOC-suspected families.

Material & Methods

Nomenclature. Nucleotide numbering is based on the cDNA sequence of *BRCA2* (NM_000059.3), c.1 denoting the first nucleotide of the translation initiation codon, as recommended by the Human Genome Variation Society.

Selection of variants predicted to affect *BRCA2* exon 3 splicing. We first selected all nucleotide changes mapping to *BRCA2* exon 3 and its flanking intronic sequences (c.68-10 to 316+40) identified in HBOC or Fanconi Anemia patient and reported in the BRCA Share™ database (Béroud *et al.*, 2016; Caputo *et al.*, 2012) (Supplementary Table 1). In addition, we selected translationally silent variations (synonymous and intronic) potentially affecting BRCA2 exon 3 splicing based on splicing-dedicated *in silico* analyses.

Splicing-dedicated *in silico* analyses. The potential impact on splicing of all translationally silent variations within *BRCA2* exon 3 (n=50) and its flanking introns (c.68-10 to 68+10, n=120) was evaluated by using two types of bioinformatics methods, splice site (ss)- and splicing regulatory elements (SRE)- dedicated methods, depending on the position of the variants relative to the exon (Supplementary Table 1).

More specifically, for intronic variants and for those mapping at exon termini on positions overlapping the splice sites, we resorted to MaxEntScan (Yeo and Burge, 2004) (MES, http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html; MaximumEntropy Model), SpliceSiteFinder-like (Shapiro and Senapathy, 1987) (SSFL, Interactive Biosoftware, Rouen). MES and SSFL were used in combined sequential fashion (MES+SSFL) as recommended by Houdayer and colleagues (Houdayer *et al.*, 2012), i.e. variants were predicted to alter splicing if MES ≤ -15% and SFFL ≤ 5%. MES and SSFL algorithms were interrogated by using the integrated software tool Alamut Batch version 1.9 (Interactive Biosoftware, Rouen, France).

For the prediction of variant-induced impact on exonic splicing regulatory elements, we resorted to four SRE-dedicated *in silico* approaches, as recently described (Tubeuf *et al.*, in preparation): (i) the QUEPASA-derived method (Ke *et al.*, 2011), which is based in the calculation of total ESR_{seq} score changes ($\Delta tESR_{seq}$) (Di Giacomo *et al.*, 2013), (ii) the HEXplorer method which calculates ΔHZ_{EI} values (Erkelenz *et al.*, 2014), (iii) the SPANR approach described by Xiong and co-workers, which yields $\Delta\Psi$ scores (Xiong *et al.*, 2015), and (iv) HAL based on the calculation of $\Delta\Psi$ scores (Rosenberg *et al.*, 2015). QUEPASA, HEXplorer, SPANR and HAL were used as stand-alone and in a combined sequential fashion (QUEPASA&HAL, AT LEAST 3 and LR_{skip}) as recommended by Tubeuf and colleagues (Tubeuf *et al.*, in preparationj), i.e. exonic variants located outside splice sites were predicted to induce exon skipping if $\Delta tESR_{seq} \leq -0.5$, $\Delta HZ_{EI} \leq -14$, $\Delta\Psi_{SPANR} \leq -0.1\%$, $\Delta\Psi_{HAL} \leq -3.4\%$ and $P_{skip} \geq 31.1\%$. Both $\Delta tESR_{seq}$ and ΔHZ_{EI} scores were calculated by using the Alamut Batch prototype tool version 1.5.2 (ESR_{seq}), (Interactive Biosoftware, Rouen, France), whereas SPANR and HAL scores were obtained by using dedicated web tools (http://tools.genes. toronto.edu and http://splicing.cs.washington. edu/ SE, respectively).

Splicing minigene reporter assays. In order to evaluate the impact on RNA splicing of the selected *BRCA2* exon 3 variants, we performed a functional assay based on the comparative analysis of the splicing pattern of wild-type (WT) and mutant reporter minigenes, as follows. Minigenes were prepared by using the pCAS2 vector (Soukarieh *et al.*, 2016), ensuing previously described procedures (Gaildrat *et al.*, 2010) with a few modifications. The WT genomic fragments ontaining BRCA2 exon 3 and 202 bp and 225 bp of upstream and downstream intronic sequence (c.68-202_ 316+225) were PCR amplified from patient genomic DNA by using a combination of forward B2Ex3_InFus_Bam-F and reverse B2Ex3_InFus_Mlu-R primers (Supplementary Table

2) and then inserted into the BamHI and MluI cloning sites of the reporter plasmid pCAS2, yielding the three-exon hybrid minigenes pCAS2-BRCA2e3 (Figure S1). When patient genomic DNA was not available, the variants of interest were introduced by site-directed mutagenesis by using the two-stage overlap extension PCR method (Ho *et al.*, 1989) and a combination of specific primers indicated in Supplementary Table 2. Then, the mutant amplicons were introduced by homologous recombination using the SLICE method (Motohashi, 2015) into the pCAS2 vector previously digested with BamHI and MluI. All constructs were sequenced to ensure that no unwanted mutations had been introduced into the inserted fragments during PCR or cloning. Next, WT and mutant minigenes (400 ng/well) were transfected in parallel into HeLa cells grown at ~70% confluence in 12-well plates using the FuGENE 6 transfection reagent (Roche Applied Science). HeLa cells obtained from ATCC were cultivated in Dulbecco's modified Eagle medium (Life Technologies) supplemented with 10% fetal calf serum in a 5% CO2 atmosphere at 37°C. Twenty-four hours later, total RNA was extracted using the NucleoSpin RNA II kit (Macherey Nagel) according to the manufacturer's instructions, and the minigenes' transcripts were analysed by semi-quantitative fluorescent RT-PCR (30 cycles of amplification) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng total RNA and the 6FAM-pCASKO1F and pCAS-2R minigene primers (Supplementary Table 2). RT-PCR products were separated by electrophoresis on 2.5% agarose gels containing ethidium bromide and visualized by exposure to ultraviolet light under saturating conditions using the Gel Doc XR image acquisition system (Bio-RAD), followed by gel-purification and sanger sequencing for proper identification of the minigene's transcripts. In parallel, splicing events were quantitated by performing capillary electrophoresis on an automated sequencer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) followed by a computational analysis with the GeneMapper v5.0 software (Applied Biosystems).

Analysis of the *BRCA2* exon 3 splicing pattern in RNA samples from patient and control individuals. Peripheral blood samples were directly collected into PAXgene Blood RNA Tubes (Qiagen) from which total RNA was extracted by using the PAXgene Blood RNA kit, according to the manufacturer's instructions. EBV-immortalized lymphoblastoid cell lines (LCLs) were cultivated in RPMI medium (Life Technologies) supplemented with 2 mM of L-glutamine and 10% fetal calf serum, at 37°C in a 5% CO2 atmosphere. Before RNA extraction, LCLs were transferred into 6-well plates, at 2.5x106 cells/well, and incubated for 5.5 hours with/without 200 µg/ml puromycin prior to harvest. Then, total RNA was extracted by using the NucleoSpin RNA

II kit (Macherey Nagel). The splicing pattern of *BRCA2* transcripts expressed in peripheral blood and in LCLs was analyzed by semi-quantitative fluorescent RT-PCR (40 and 26 cycles of amplification for PAXgene and LCL samples, respectively) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng of total RNA and a combination of forward and reverse primers located in *BRCA2* exons 2 and exon 5, respectively (Supplementary Table 2). Then, RT-PCR products were separated by electrophoresis on a 2% agarose gel, gel-purified and sequenced. In parallel, splicing events were quantitated by performing capillary electrophoresis on an automated sequencer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) and analyzed by using the GeneMapper v5.0 software (Applied Biosystems).

Allele-specific expression analysis. Allele-specific expression (ASE) was measured by performing a SNaPshot quantitative primer extension assay (SNaPshot MultiplexKit, Applied Biosystem), as previously described in Soukarieh *et al.*, 2016. Briefly, RT-PCR products spanning *BRCA2* exons 2 to 5 were obtained from patients' RNA sample (PAXgenes and LCLs), by using a combination of forward RT-BRCA2Ex2-F and reverse RT-BRCA2Ex5-R primers (Supplementary Table 2). In parallel, the genomic segment encompassing *BRCA2* exon 3 was amplified by PCR from the genomic DNA of the same patient by using a combination of forward BRCA2Ex3_InFus_Bam-F and reverse BRCA2Ex3_InFus_Mlu-R (Supplementary Table 2). Then, primer extension reactions were performed by using the RT-PCR and PCR products as template and variant-specific primers indicated in Supplementary Table S2. Purified extension products along with 120 LIZ Size Standard (Applied Biosystems) were separated by capillary electrophoresis on an automated 3500 Genetic Analyzer (Applied Biosystems) and analyzed by using the GeneMapper v5.0 software (Applied Biosystems). SNaPshot results obtained from patient cDNA were normalized by those obtained from patient gDNA.

Mouse embryonic stem cells (mESC)-based complementation assay. PL2F7 cells were generated from AB2.2 mouse embryonic stem cell line by knocking out one copy of Brca2 and flanking the other allele of Brca2 with two LoxP sites (Kuznetsov *et al.*, 2008, 2010). PL2F7 were cultured on mitotically inactive SNL feeder cells in M15 media, which is Knockout DMEM media (Life Technologies) supplemented with 15% fetal bovine serum (FBS; Life Technologies), 0.00072% beta-mercaptoethanol, 100 U ml 1 penicillin, 100 mg ml 1 streptomycin and 0.292 mg ml 1 L-glutamine at 37 C, in a 5% CO_2 atmosphere. Genomic fragments containing the variant of

interest were PCR-amplified from mutant pCAS2-BRCA2e3 minigenes by using specific primers carrying BAC homology arms (Supplementary Table 2). These fragments were subsequently introduced into BAC constructs containing the full-length human *BRCA2* gene (BAC RPCI-11 777 19I) by recombineering in SW102 bacteria using the galK-based counter selection method as previously described (Chang *et al.*, 2012; Kuznetsov *et al.*, 2010; Sharan *et al.*, 2009). BAC DNA (25 µg) was then electroporated into 1.10^7 mESC suspended in 0.9 ml PBS by setting a Gene Pulser (Bio-Rad) at 230 V, 500 mF. Thirty six hours after electroporation, G418 (180µg/mL) selection was performed for 5 days, after which cells were transferred to normal M15 medium until colonies became visible. Forty eight individual colonies were picked into 96-well plates and hBRCA2 expression was verified by RT-PCR and Western blot analyses. For RT-PCR analysis, total RNA was extracted from mESC pellets by using RNeasy 96 kit (Qiagen), according to the manufacturer instruction and RT-PCR (30 cycles of amplification) was carried out in a 25µL reaction volume by using the OneStep RT-PCR kit (ABM), ~100 ng total RNA and B2ex11FRT and B2ex14RRT primers, located in exons 11 and 14, respectively (Supplementary Table S1). RT-PCR products were separated by electrophoresis on 1.2% agarose gels containing ethidium bromide and visualized by exposure to ultraviolet light. For Western blot, cells were lysed in lysis buffer (20 mM HEPES, 100 mM NaCl, 1mM EDTA, 1mM NaF, 1mM EGTA, 1 mM DTT and 0,1% Triton X) and protein extracts were resolved by SDS PAGE by using NuPAGE^{TM} 3-8% Tris-Acetate protein gels (Invitrogen) and subsequently transferred to nitrocellulose membranes. Blots were blocked with milk and incubated with the rabbit polyclonal anti-BRCA2 (BETHYL, A303-434A-T-1) primary antibody overnight at 4°C, washed with PBST and probed with horseradish peroxidase-conjugated secondary antibodies at room temperature for 2 h and subjected to ECL (Amersham). Immunoreactive bonds were detected by immunoblots incubated with ECL reagent and exposed for 30 sec to 10 min in a luminescence image analyser. After confirmation of hBRCA2 expression, $Brca2^{cko}$ allele was deleted by electroporation of PGK-Cre plasmid DNA (25 µg) into 1.10^7 mESC $G418^R$ suspended in 0.9 ml PBS by setting a Gene Pulser (Bio-Rad) at 230 V, 500 mF. 36 hr after electroporation, HAT (Hypoxanthine-Aminopterin-Thymidine) selection of $Brca2^{ko/ko}$ mESC was performed for 5 days, after which cells were switched to HT (Hypoxanthine-Thymidine) media for 2 days and then transferred to normal M15 medium until colonies became visible. HAT resistant (HAT^R) colonies were counted after methylene blue staining by using ImageJ and the number of colonies was then compared to that of no HAT control to determine the

rescue rate (HATx100/noHAT). In parallel, 24 individual colonies were picked into 96-well plates and were genotyped by southern blotting as previous described (Ding *et al.*, 2016; Kuznetsov *et al.*, 2010). After confirmation of BRCA2 recue, two HAT^R clones were selected for genotoxin and irradiation sensitivity assays. 8,000 mESC HAT^R per well were seeded in triplicates in 96-well plates. Eighteen hours after seeding, drug treatment (Cisplatin, Mitomycin C (MMC), Methyl methanesulfonate (MMS), Olaparib and Camptothesin) was performed for 72 hours, whereas 24 hours after seeding, ionizing irradiation (^{137}Cs source, γ-irradiator) was performed and then cells were cultured in fresh M15 media for 72 hours, after which the relative number of living cells was measured using XTT assay (Kuznetsov *et al.*, 2010; Scudiero *et al.*, 1988). Drug treatment concentrations as well as irradiation dose exposition are indicated in Figure 3. In addition, the splicing pattern of hBRCA2 transcripts expressed in HAT^r mES cells was analyzed by semi-quantitative fluorescent RT-PCR (26 cycles of amplification) in a 25 µl reaction volume by using the OneStep RT-PCR kit (Qiagen), 200 ng of total RNA and a combination of specific primers located in *BRCA2* exons 2 and exon 5 (Table S1). Then, RT-PCR products were separated by electrophoresis on a 2% agarose gel, gel-purified and sequenced and splicing events were quantitated by performing capillary electrophoresis on an automated sequencer (Applied Biosystems) using 500 ROX™ Size Standard (Applied Biosystems) and computational analysis by using the GeneMapper v5.0 software (Applied Biosystems).

Performance assessment. The evaluation of the predictive power of splicing-dedicated bioinformatics methods was performed by measuring sensitivity (Sen) = [TPx100/(TP + FN)], specificity (Sp) = [TNx100/(TN + FP)], accuracy (Acc) = [(TN+TP)x100/(TN+TP+FN+FP), where TP (true positive) and FN (false negative) values are the numbers of positive samples that are predicted to be positive and negative respectively. Analogously, TN (true negative) and FP (false positive) values are the numbers of negative samples that are predicted to be negative and positive respectively. TP, TN, FP, FN were determined by taking into account thresholds determined either previously (Tubeuf *et al.*, in preparation), as indicated, or by performing new ROC (Receiver operating characteristic) curves representing the "closest to top left" (i.e. by minimizing the distance at the top left corner). ROC curves were performed by using GraphPad Prism software (Version 5.0) and easyROC (http://www.biosoft.hacettepe.edu.tr/easyROC/). The predictive power of the SRE-dedicated tools were then compared to each other by using Venn

diagrams plotted by Jvenn (Bardou *et al.*, 2014), an interactive web application (http://jvenn.toulouse.inra.fr/app/ example.html).

Results

Selection of 101 variants within *BRCA2* exon 3 and its flanking intronic regions.

A recent collaborative study conducted by the French COVAR clinical trial group and the ENIGMA consortium reported the pathogenicity of complete loss of *BRCA2* exon 3 by conducted multifactorial likelihood analyses (Caputo *et al.*, 2018). Particularly, this study focused on 8 *BRCA2* variants resulting in complete deletion of exon 3 at the RNA level (Δ3) demonstrate that variants leading to a total skipping of exon 3 are associated with an increased risk of breast and ovarian cancer (Caputo *et al.*, 2018). However, contrary to variant-induced total exon 3 skipping, the impact of variants leading to partial exon 3 skipping on BRCA2 function remains unknown and the pathogenicity of such variants have yet to be established.

In order to better understand the contribution to pathogenicity of variations that affect *BRCA2* exon 3 splicing, in particular those inducing partial splicing defects, we analysed the entire set of nucleotide variations (n=50) identified within *BRCA2* exon 3 and the 25 bp franking intronic regions (c.68-8 to c.631+50) and reported within the BRCA-Share database by the French BRCA diagnostic laboratories (Supplementary Table 1). We also included in our cohort the c.316G>A variation identified in a child who was born alive but die of malignancy associated with Fanconi anemia. In addition, to avoid bias of change in the amino acid sequence of the protein in multifactorial likelihood analyses, we also selected additional translationally silent variations (intronic and synonymous variants) mapping to *BRCA2* exon 3 and the flanking intronic regions (c.68-10 to c.631+10) potentially affecting *BRCA2* exon 3 splicing (n=51, Supplementary Table 1), either by disrupting splice site (n=13, Supplementary Table 3) or splicing regulatory elements (n=38, Supplementary Table 4) based on slicing-dedicated *in silico* tools.

As a result, we collected a total of 101 newly reported natural variations, including 95 SNVs and 6 small deletions, most of which (n=67) identified in cancer patients suspected of HBOC (Supplementary Table 1). Only 10 of these variations are currently classified as clearly pathogenic and 8 as clearly not pathogenic, whereas the remaining 49 natural variations include either variations not yet classified (n=3), VUS (n=18) or variations with conflicting interpretation (n=28)

(Supplementary Table 1). Altogether, this selection comprised 24 intronic variations and 76 exonic SNVs, which were assessed for their impact on splicing by using functional splicing assays.

Identification of an unexpected high proportion of mutations affecting the splicing pattern of *BRCA2* exon 3 by using a splicing minigene-based assay.

We performed an *ex vivo* splicing assay based on pCAS2-*BRCA2e3*-derived minigenes to assess the impact of 101 selected variants on *BRCA2* exon 3 splicing (Figure S1). Importantly, the minigene assay revealed that 65 out of the 101 variations (64%) altered the splicing pattern of exon 3 relative to wild-type whereas the 36 remaining variations showed no effect on splicing (36%) (Supplementary Table 1). More precisely, 63 variations induced exon 3 skipping (61%) to different extents, 1 variation (c.68-1G>A) is responsible of the complete deletion of six nucleotides at the start of the exon 3 due to the creation of an acceptor site at the exonic position +6 (Figure S2A) and 1 variation caused 2 partial RNA splicing defects, i.e. skipping of exon 3 (Δ3 = 14%) and the 45-nucleotide deletion at the beginning of the exon (9%), due to the creation of an acceptor site at the exonic position +44, that compete with the reference 3'ss (Figure S2B & C) according to splice site-dedicated algorithms.

We surmise, given their position, that 22 out of the 65 (34%) splicing mutations detected in exon 3 directly affect the definition of the reference splice sites, either at the level of the 3'splice site (n=5, Supplementary Table 3) or of the 5'splice site (n=16, Supplementary Table 4) (Cartegni *et al.*, 2002).The effects produced by most of these 22 variants could have been correctly predicted by algorithms commonly used to predict alterations of reference splice sites, such as Splice Site Finder Like (SSFL, n=19) and MaxEntScan (MES, n=15) and their combination (MES+SSFL = 13) (Supplementary Table 3). Interestingly, SPiCE (Splicing Prediction in Consensus Elements), a newly developed prediction tool that uses logistic regression to combine *in silico* predictions from SSFL and MES, correctly predicted the effects induced by the quasi totality variations impairing splice sites (21 out of 22, 96% accuracy, Supplementary Table 3). Because the 43 remaining splicing mutations detected in *BRCA2* exon 3 map outside the positions directly defining the splice sites, we strongly suspect that they interfere with exon recognition by altering ESRs. Accordingly, the effects produced of most of the 43 variants could have been correctly predicted by using SRE-dedicated *in silico* tools with previously established generic thresholds (Tubeuf *et al.*, in

preparation). More concretely, as shown in Supplementary Table 4, 28 to 38 out of 43 variants, depending on the SRE-dedicated *in silico* tools were correctly predicted (65%-88% range in sensitivity), confirming the reliability of SRE-dedicated *in silico* approaches in pinpointing variant-induced ESR alterations. However, we observed an important rate of false positive calls (32%-81% range in specificity), suggesting that these approaches may need further adjustment to be appliqued successfully on *BRCA2* exon 3. Altogether, our data revealed that a striking high proportion of variations affect the splicing pattern of *BRCA2* exon 3 (64%) and indicate an important contribution of variant-induced ESR alterations, suggesting that this exon is particularly sensitive to ESR-mutation.

Confirmation of variant spliceogenic effect from RNA samples of HBOC patients.

Although a high concordance for splicing analysis was demonstrated for multiple disease-associated *BRCA1/2* variants between splicing reporter minigene assays and blood-derived mRNA samples, differences in splicing pattern have been observed for a proportion of spliceogenic variants (Acedo *et al.*, 2012; Bonnet *et al.*, 2008; van der Klift *et al.*, 2015; Steffensen *et al.*, 2014). To apprehend the physiological relevance of the splicing defects revealed in the splicing reporter minigene assay, we then decided to compare our results with data derived from the analysis of *BRCA2* exon 3 splicing pattern in blood-derived RNA samples obtained from carriers of variants at the heterozygous state. In collaboration with the BRCA diagnostic laboratories within the Genetics and Cancer Group (GGC), we had the opportunity to collect 18 RNA samples for 11 SNVs from either PAXgene-stabilized blood (PAXgene, n=2) and lymphoblastoide cell lines (LCLs, n=9). In addition, we were able to obtain a LCL RNA samples for one patient harboring an Alu insertion in the middle of exon 3 (c.156_157insAlu) associated to total skipping of *BRCA2* exon 3 (Machado *et al.*, 2007) and used as control in our analysis.

RT-PCR characterisation of variant-induced BRCA2 exon 7 splicing defects in blood-derived samples. Analysis of 18 control RNA samples (n=3 and 15 for PAXgenes and LCLs, respectively) and by semi-quantitative fluorescent RT-PCR using primers targeting exons 2 and 5 revealed the production of two transcripts: one major transcript corresponding to normal exon 3 inclusion (+E3 = 89% and 92% in peripheral blood and LCL-derived RNA samples, respectively) and one minor transcript without the exon 3 (Δ3 = 10 and 8%) (Supplementary Table 1). These data support the existence of an alternative exon 3 splicing, in good agreement with previously published data

(Davy *et al.*, 2017; Fackenthal *et al.*, 2016). In addition, analysis of the splicing pattern of *BRCA2* exon 3 in LCL-derived RNA sample from the individual harbouring the c.156_157insAlu at the heterozygous state revealed a major increase in the relative levels of transcripts without exon 3, as expected (Δ3 = 70%), as compared to ~8% in controls LCLs (Supplementary Table 1) (Machado *et al.*, 2007) . We then analysed the splicing pattern of *BRCA2* exon 3 in patient biological samples and compared to those generated from equivalent RNA samples of the healthy control individuals. Results shown in Supplementary Table 1 indicates that only the patient carrying the c.223G>C variation had a splicing pattern similar to that of healthy controls, in agreement with the minigene results and confirming that this variation did not affect exon 7 splicing *in vivo*. In contrast, RNA from patients harboring either c.68-8_68-7delinsAA, c.68-7T>A, c.92G>A, c.102A>G, c.145G>T, c.231T>G, c.289G>T, c.316G>A, c.316G>C or c.316+6T>C are associated with a decrease in the amount of FL *BRCA2* transcripts and a concomitant increase in transcripts lacking exon 3 ($25\% \leq \Delta 3 \leq 63\%$), as compared to healthy controls (Δ3= 8-11%), indicating that those variants alter the splicing pattern of *BRCA2* exon 3. Nevertheless, the level of exon 3 skipping is lower than that observed in the patient harboring the alu insertion (Δ3 = 70%), suggesting that those variants do not induce a total splicing defect. Sequencing of the FL RT-PCR products of patients carrying the exonic variants showed the presence of both WT and mutant FL transcripts, but mutant FL transcripts seemed to be less represented in lower proportion compared to WT FL transcripts, further indicating that those variations cause weak to severe partial splicing defects (Figure S3).

Allele specific expression analyses. In order to better evaluate the contribution of wild-type and mutant alleles to the production of *BRCA2* transcripts containing exon 3, we took advantage of the quantitative nature of SNaPshot assay allowing to measure allele specific expression (ASE), i.e. the relative amount of each allele within the RT-PCR product containing exon 3. As expected, we do not observed an allelic imbalance for the c.223G>C variation (ASE = 96% of WT), which confirms that this variation does not affect exon 3 splicing (Supplementary Table 1). In contrast, analysis of the bi-allelic expression indicates that the FL transcripts expressed from the mutant allele were in fact present in the cells of patients carrying the c.92>A (ASE = 48%), c.102A>G (ASE = 53%), c.145G>T (ASE = 49%), c.231T>G (ASE = 55-65%, depending on the patient), c.289G>T (ASE = 52%), c.316G>A (ASE = 23%) and c.316G>C (ASE = 39%) variations (Supplementary Table 1). Given the results of our minigene assays and RT-PCR analysis of patient's RNA, we conclude that this allelic imbalance is essentially due to variant-induced exon 3

skipping and that these variations induce partial exon skipping, but not total splicing defects. Surprisingly, we did not observe an allelic imbalance for one of the patient carrying the c.231T>G variation (ASE = 108%) despite this variation increase exon 3 skipping in the minigene assay and RT-PCR analysis of patient's RNA as well (Supplementary Table 1), suggesting that another variation located on the alleged WT allele might be responsible of a decrease of transcripts generated from the alleged WT allele. Indeed, the patient carrying this c.231T>G variation also harbour, in trans, a deleterious variation located in exon 11, c.6515C>A expected to lead to the introduction of a premature termination codon (PTC) at the 2172th amino acid (p.(Ser2172*)), which is most likely targeted by the NMD decay.

Comparison of patient RNA data with results from the ex vivo splicing reporter minigene-based assay. Altogether, the in vivo results agree with the pCAS2 minigene assays and highlight the physiological pertinence of the pCAS2-BRCA2e3 minigene-based splicing assay. Nonetheless, it is important to note that the WT pCAS2-*BRCA2*e3 minigene did not fully reproduce the splicing pattern of the endogenous *BRCA2* exon 3 expressed in the control RNA samples, as the alternative skipping of *BRCA2* exon 3 observed in patient's RNA (Δ3 = 8-11%) is much lower in minigene assay (Δ3 = 1%), suggesting that the minigene system might underestimate the severity of the observed splicing defects. However, we observed a logarithmic correlation (R^2 = 0.9819) between the severity of the splicing defects evaluated in the monoallelic minigene-based assay and those assessed in LCLs derived from patients carrying the same variant at the heterozygous state (Figure S4).

Evaluation of consequences on protein function of partial *BRCA2* exon 3 skipping by using a mouse embryonic stem cell (mESC)-based functional assay.

Among the 67 splicing mutations identified by our RNA splicing assays, only 18 (1 missense, 4 nonsense, 1 frame-shift and 12 intronic) are currently classified as pathogenic (Supplementary Table 1), including variations that alter the open reading frame and variations that lead to total exon 3 skipping, recently classified as pathogenic on the basis of co-segregation analysis and functional data (Biswas *et al.*, 2012; Caputo *et al.*, 2018). In contrast, most of the variations associated with a partial exon 3 skipping are considered as VUS.

***Selection of* tr*anslationally silent VUS causing increasing levels of exon 3 skipping*.** In order to better understand the contribution to pathogenicity of variant-induced partial *BRCA2* exon 3 skipping, we selected translationally silent variations (either intronic or synonymous) causing increasing levels of exon skipping (gradient) according to the results generated from functional splicing reporter minigene-based assay (Supplementary Figure 5). As a result, 9 variations (8 natural and 1 artificial) causing increasing levels of exon 3 skipping from 7% to 95% were retrained for downstream analysis (Figure 1). This selection encompassed the two control variations c.68-7T>A (Δ3 = 7%) and c.316+5G>C (Δ3 = 95%), currently classified as neutral and deleterious, respectively, based on clinical, genetic and familial data (Caputo *et al.*, 2018; Colombo *et al.*, 2018) and 7 variations currently considered as VUS (Δ3 = 15% to 77%) (Figure 1 and Supplementary Figure 5). In addition, we also selected a missense variation c.316G>A (p.Gly106Arg) which results in partial exon 3 skipping (Δ3 = 65%) identified in trans with a deleterious variant c.6515C>A (p.Ser2172*) in a child with FA who passed away at age 2, suggesting that the c.316G>A variation might encode a partially functional BRCA2 protein allowing foetal viability. To validate the pertinence of our selection based on the results obtained with pCAS2-BRCA2e3 minigenes, we then decided to verify if the splicing defects could also be detected in a physiological context. As shown on Figure 1, the severity of all the splicing defects observed in the minigene assay was confirmed in patient blood-derived samples.

We then endeavour to evaluate the consequences of variant-induced partial *BRCA2* exon 3 skipping on BRCA2 protein function by performing a more comprehensive functional assay aiming to evaluate the consequences of the combined effect of any BRCA2 variant on RNA splicing and protein function, in a genomic, biological and cellular context close to the natural context. To this end, we took advantage of mouse embryonic stem cells (mESC)-based complementation assay to evaluate the ability of cis-acting variations to rescue the lethal cell phenotype of loss of *BRCA*2 function after Cre-mediated removal of the conditional *BRCA*2 allele (Kuznetsov *et al.*, 2008, 2010). Nonfunctional protein variants that fail to rescue the ES cell lethality are considered pathogenic whereas variants resulting in functional BRCA2 protein that are able to fully or partially overcome Brca2 loss are likely to be neutral or pathogenic, depending on their capacity to perform BRCA2 functions (Kuznetsov *et al.*, 2008; Mesman *et al.*, 2018).

Comparisons of the RT- PCR analyses of in vivo mESC RNA with data from the ex vivo splicing reporter minigene-based assay. To date, a very small number of BRCA2 variants has been analysed in the mESC-based functional assay, especially those affecting mRNA splicing. Still, the mESC-based system looks very promising as it allows the monoallelic analysis of the functional impact of any BRCA2 variant, including those located in either exonic or intronic sequences that may affect RNA splicing. To validate the physiological pertinence of such system for the evaluation of variant-induced BRCA2 exon 3 splicing defects, we compared the results from the pCAS2 minigene-based splicing assay with the results from the HAT^r mESC-derived mRNA samples, both monoallelique splicing assays (Figure 1). For the 8 variations in addition to the WT, the results of the HAT^r mESC-derived mRNA analysis were fully consistent with those of the minigene assay not only in terms of whether or not a variant caused aberrant splicing but also in terms of the severity of the splicing defects (Figure 1). Indeed, we found a linear correlation (R^2 = 0.9973) between the level of exon skipping assessed in the minigene-based splicing assay and those measured in the HAT^r mESC carrying the same variant (Supplementary Figure 4). Moreover, it is important to note that the WT HATr mESC (Δ3 = 10%) fully reproduce the splicing pattern of the endogenous *BRCA2* exon 3 expressed in the healthy individual mRNA samples (Δ3 = 8%) (Figure 1). Altogether, the results obtained in HAT^r mESC agree with those obtained in patient-derived RNA samples and highlight the physiological pertinence of the mESC model to evaluate the consequence on splicing of BRCA2 exon 3 VUS.

Rescue of mouse Brca2$^{KO/KO}$ ES cells by human BRCA2 variants. Functional BRCA2 is essential for cell survival and normal embryonic development in mice (Hakem *et al.*, 1998; Sharan *et al.*, 1997). We first tested the ability of the class 1 (c.68-7T>A) and class 5 (c.316+5G>C) variations to rescue the lethal cell phenotype due to loss of BRCA2 function after Cre-mediated removal of the conditional *BRCA2* WT allele. Only the cells carrying the c.316+5G>C variation were unable to form HAT-resistant (HAT^r) colonies in the absence of functional BRCA2, indicating that BRCA2 function was severely impaired and supporting the deleterious nature of this variant (Figure 2 and Table 1). In contrast, the considered neutral variation (c.68-7T>A) was able to fully complement loss of endogenous Brca2 (Figure 2 and Table 1). We then expanded our analysis to the VUS of interest inducing increased levels of exon 3 skipping. The complementation phenotype of 5 out of the 8 VUS tested, i.e., variants c.165C>T, c.231T>G, c.102A>G, c.68-8_-7delinsAA and c.316+6T>C resembled that of the class 1 variation with respect to their full ability to rescue

the cell lethality imposed by Cre-mediated loss of *BRCA2* (Figure 2 and Table 1). In the case of variations c.316+6T>G, c.316+6T>A and c.316G>A, we observed smaller numbers of HAT-resistant clones after removal of the conditional *BRCA2* allele, as compared to WT *BRCA2*-expressing cells (Figure 2 and Table 1), suggesting that BRCA2 function is compromised in these cells resulting in incomplete complementation.

Sensitivity of HAT^r mES cells to genotoxins and irradiation. Because of the known role of the repair of DSBs, the absence of functional BRCA2 protein will render cells vulnerable to compounds that introduce toxic DNA lesions that impede cellular processes such as transcription and replication. Therefore, we next tested the sensitivity of surviving ES-cells (HAT^r) to genotoxins (Cisplatin, Mitomycin C, Methyl methanesulfonate, Olaparib and Camptothesin) and irradiation by cell survival measurements. As expected, the *BRCA2* c.68-7T>A-expressing HATr cells exhibited no difference in sensitivity to various DNA damaging agents compared to control cells (Table 1). Similarly, cells carrying the c.165C>T, c.231T>G, c.102A>G, c.68-8_-7delinsAA and c.316+6T>C did not exhibit hypersensitivity to none of the DNA-damaging agents strongly supporting their neutrality (Figure 3 and Table 1). In contrast, c.316+6T>G, c.316+6T>A and c.316G>A are associated with a severe hypersensitivity to various DNA damaging agents compared to WT or c.68-7T>A *BRCA2*-expressing cells, strongly suggesting an impairment of BRCA2 function (Figure 3 and Table 1).

Association of partial BRCA2 exon 3 skipping with breast and/or ovarian cancer risk by multifactorial likelihood analyses (Sandrine Caputo, in progress).

Complementary multifactorial likelihood analyses, based on the collect of genetic, clinical, tumoral, cosegregation and familial data of patients carrying the natural variations of interest, are currently underway in order to further evaluate the pathogenic nature of this type of variation.

Bibliography

Acedo, A., Sanz, D.J., Durán, M., Infante, M., Pérez-Cabornero, L., Miner, C., and Velasco, E.A. (2012). Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. Breast Cancer Res. BCR *14*, R87.

Ars, E., Serra, E., García, J., Kruyer, H., Gaona, A., Lázaro, C., and Estivill, X. (2000). Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum. Mol. Genet. *9*, 237–247.

Baralle, D., Lucassen, A., and Buratti, E. (2009). Missed threads. The impact of pre-mRNA splicing defects on clinical practice. EMBO Rep. *10*, 810–816.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. BMC Bioinformatics *15*, 293.

Béroud, C., Letovsky, S.I., Braastad, C.D., Caputo, S.M., Beaudoux, O., Bignon, Y.J., Bressac-De Paillerets, B., Bronner, M., Buell, C.M., Collod-Béroud, G., *et al.* (2016). BRCA Share: A Collection of Clinical BRCA Gene Variants. Hum. Mutat. *37*, 1318–1328.

Biswas, K., Das, R., Eggington, J.M., Qiao, H., North, S.L., Stauffer, S., Burkett, S.S., Martin, B.K., Southon, E., Sizemore, S.C., *et al.* (2012). Functional evaluation of BRCA2 variants mapping to the PALB2-binding and C-terminal DNA-binding domains using a mouse ES cell-based assay. Hum. Mol. Genet. *21*, 3993–4006.

Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., Berthet, P., Chevrier, A., Dugast, C., Layet, V., *et al.* (2008). 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. *45*, 438–446.

Caputo, S., Benboudjema, L., Sinilnikova, O., Rouleau, E., Béroud, C., Lidereau, R., and French BRCA GGC Consortium (2012). Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. Nucleic Acids Res. *40*, D992-1002.

Caputo, S.M., Léone, M., Damiola, F., Ehlen, A., Carreira, A., Gaidrat, P., Martins, A., Brandão, R.D., Peixoto, A., Vega, A., *et al.* (2018). Full in-frame exon 3 skipping of BRCA2 confers high risk of breast and/or ovarian cancer. Oncotarget *9*, 17334–17348.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Chang, S., Stauffer, S., and Sharan, S.K. (2012). Using recombineering to generate point mutations: the oligonucleotide-based "hit and fix" method. Methods Mol. Biol. Clifton NJ *852*, 111–120.

Colombo, M., Lòpez-Perolio, I., Meeks, H.D., Caleca, L., Parsons, M.T., Li, H., De Vecchi, G., Tudini, E., Foglia, C., Mondini, P., *et al.* (2018). The BRCA2 c.68-7T > A variant is not pathogenic: A model for clinical calibration of spliceogenicity. Hum. Mutat. *39*, 729–741.

Davy, G., Rousselin, A., Goardon, N., Castéra, L., Harter, V., Legros, A., Muller, E., Fouillet, R., Brault, B., Smirnova, A.S., *et al.* (2017). Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. Eur. J. Hum. Genet. EJHG *25*, 1147–1154.

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., and Martins, A. (2013). 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. *34*, 1547–1557.

Ding, X., Ray Chaudhuri, A., Callen, E., Pang, Y., Biswas, K., Klarmann, K.D., Martin, B.K., Burkett, S., Cleveland, L., Stauffer, S., *et al.* (2016). Synthetic viability by BRCA2 and PARP1/ARTD1 deficiencies. Nat. Commun. *7*, 12425.

Eccles, D.M., Mitchell, G., Monteiro, A.N.A., Schmutzler, R., Couch, F.J., Spurdle, A.B., Gómez-García, E.B., and ENIGMA Clinical Working Group (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. *26*, 2057–2065.

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J.O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res. *42*, 10681–10697.

Fackenthal, J.D., Yoshimatsu, T., Zhang, B., de Garibay, G.R., Colombo, M., De Vecchi, G., Ayoub, S.C., Lal, K., Olopade, O.I., Vega, A., *et al.* (2016). Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples. J. Med. Genet. *53*, 548–558.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. Clifton NJ *653*, 249–257.

Gaildrat, P., Krieger, S., Di Giacomo, D., Abdat, J., Révillion, F., Caputo, S., Vaur, D., Jamard, E., Bohers, E., Ledemeney, D., *et al.* (2012). Multiple sequence variants of BRCA2 exon 7 alter splicing regulation. J. Med. Genet. *49*, 609–617.

Hakem, R., de la Pompa, J.L., and Mak, T.W. (1998). Developmental studies of Brca1 and Brca2 knock-out mice. J. Mammary Gland Biol. Neoplasia *3*, 431–445.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., and Pease, L.R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene *77*, 51–59.

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., *et al.* (2012). 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. *33*, 1228–1238.

Kast, K., Rhiem, K., Wappenschmidt, B., Hahnen, E., Hauke, J., Bluemcke, B., Zarghooni, V., Herold, N., Ditsch, N., Kiechle, M., *et al.* (2016). Prevalence of BRCA1/2 germline mutations in 21 401 families with breast and ovarian cancer. J. Med. Genet. *53*, 465–471.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

van der Klift, H.M., Jansen, A.M.L., van der Steenstraten, N., Bik, E.C., Tops, C.M.J., Devilee, P., and Wijnen, J.T. (2015). 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. *3*, 327–345.

Kuznetsov, S.G., Liu, P., and Sharan, S.K. (2008). Mouse embryonic stem cell-based functional assay to evaluate mutations in BRCA2. Nat. Med. *14*, 875–881.

Kuznetsov, S.G., Chang, S., and Sharan, S.K. (2010). Functional analysis of human BRCA2 variants using a mouse embryonic stem cell-based assay. Methods Mol. Biol. Clifton NJ *653*, 259–280.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980-985.

Ledermann, J., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., Scott, C.L., Meier, W., Shapira-Frommer, R., Safra, T., *et al.* (2014). Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. Lancet Oncol. *15*, 852–861.

Leman, R., Gaildrat, P., Gac, G.L., Ka, C., Fichou, Y., Audrezet, M.-P., Caux-Moncoutier, V., Caputo, S.M., Boutry-Kryza, N., Léone, M., *et al.* (2018). 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.

Lindor, N.M., Goldgar, D.E., Tavtigian, S.V., Plon, S.E., and Couch, F.J. (2013). BRCA1/2 sequence variants of uncertain significance: a primer for providers to assist in discussions and in medical management. The Oncologist *18*, 518–524.

Machado, P.M., Brandão, R.D., Cavaco, B.M., Eugénio, J., Bento, S., Nave, M., Rodrigues, P., Fernandes, A., and Vaz, F. (2007). Screening for a Rearrangement in High-Risk Breast/Ovarian Cancer Families: Evidence for a Founder Effect and Analysis of the Associated Phenotypes. J. Clin. Oncol. *25*, 2027–2034.

Martinez, J.S., Baldeyron, C., and Carreira, A. (2015). Molding BRCA2 function through its interacting partners. Cell Cycle Georget. Tex *14*, 3389–3395.

Mesman, R.L.S., Calléja, F.M.G.R., Hendriks, G., Morolli, B., Misovic, B., Devilee, P., van Asperen, C.J., Vrieling, H., and Vreeswijk, M.P.G. (2018). The functional impact of variants of uncertain significance in BRCA2. Genet. Med. Off. J. Am. Coll. Med. Genet.

Motohashi, K. (2015). A simple and efficient seamless DNA cloning method using SLiCE from Escherichia coli laboratory strains and its application to SLiP site-directed mutagenesis. BMC Biotechnol. *15*, 47.

Muller, D., Rouleau, E., Schultz, I., Caputo, S., Lefol, C., Bièche, I., Caron, O., Noguès, C., Limacher, J.M., Demange, L., *et al.* (2011). An entire exon 3 germ-line rearrangement in the

BRCA2 gene: pathogenic relevance of exon 3 deletion in breast cancer predisposition. BMC Med. Genet. *12*, 121.

Nielsen, F.C., van Overeem Hansen, T., and Sørensen, C.S. (2016). Hereditary breast and ovarian cancer: new genes in confined pathways. Nat. Rev. Cancer *16*, 599–612.

Prakash, R., Zhang, Y., Feng, W., and Jasin, M. (2015). Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. Cold Spring Harb. Perspect. Biol. *7*, a016600.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Sanz, D.J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardeñosa, E., Lastra, E., Pagani, F., Miner, C., and Velasco, E.A. (2010). A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *16*, 1957–1967.

Scudiero, D.A., Shoemaker, R.H., Paull, K.D., Monks, A., Tierney, S., Nofziger, T.H., Currens, M.J., Seniff, D., and Boyd, M.R. (1988). Evaluation of a soluble tetrazolium/formazan assay for cell growth and drug sensitivity in culture using human and other tumor cell lines. Cancer Res. *48*, 4827–4833.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155–7174.

Sharan, S.K., Morimatsu, M., Albrecht, U., Lim, D.S., Regel, E., Dinh, C., Sands, A., Eichele, G., Hasty, P., and Bradley, A. (1997). Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. Nature *386*, 804–810.

Sharan, S.K., Thomason, L.C., Kuznetsov, S.G., and Court, D.L. (2009). Recombineering: a homologous recombination-based method of genetic engineering. Nat. Protoc. *4*, 206–223.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., Tosi, M., and Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using *In silico* Tools. PLoS Genet. *12*, e1005756.

Spurdle, A.B., Healey, S., Devereau, A., Hogervorst, F.B.L., Monteiro, A.N.A., Nathanson, K.L., Radice, P., Stoppa-Lyonnet, D., Tavtigian, S., Wappenschmidt, B., *et al.* (2012). 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. *33*, 2–7.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T. vO (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. Eur. J. Hum. Genet. EJHG *22*, 1362–1368.

Szabo, C., Masiello, A., Ryan, J.F., and Brody, L.C. (2000). The breast cancer information core: database design, structure, and scope. Hum. Mutat. *16*, 123–131.

Teraoka, S.N., Telatar, M., Becker-Catania, S., Liang, T., Onengüt, S., Tolun, A., Chessa, L., Sanal, O., Bernatowska, E., Gatti, R.A., *et al.* (1999). Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am. J. Hum. Genet. *64*, 1617–1631.

Théry, J.C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M.-P., Killian, A., Duponchel, C., Rousselin, A., Vaur, D., Peyrat, J.-P., *et al.* (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur. J. Hum. Genet. *19*, 1052–1058.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J., *et al.* (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. *29*, 1412–1424.

Walker, L.C., Whiley, P.J., Houdayer, C., Hansen, T.V.O., Vega, A., Santamarina, M., Blanco, A., Fachal, L., Southey, M.C., Lafferty, A., *et al.* (2013). 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. *34*, 1424–1431.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. J. Comput. Mol. Cell Biol. *11*, 377–394.

Zou, J.P., Hirose, Y., Siddique, H., Rao, V.N., and Reddy, E.S. (1999). Structure and expression of variant BRCA2a lacking the transactivation domain. Oncol. Rep. *6*, 437–440.

Legends to figures

Table 1. Functional evaluation of *BRCA2* exon 3 variants causing increasing levels of exon 3 skipping. The impact of 9 selected variants on RNA splicing and protein function was determined.

[1] Variant classification is indicated when attributed and was retrieved from each database and refers to the 5-tier system used by the InSiGHT Variant Interpretation Committee (http://insight-group.org/variants/classifications/) as follows: 1, not pathogenic or begnign; 2, likely not pathogenic or likely begnign; 3, uncertain significance (also called VUS for variants of unknown significance); 4, likely pathogenic; 5, pathogenic; n/a, not available.

[2] For each splicing functional assay, the relative quantification of splicing events was evaluated by fluorescent RT-PCR followed by capillary electrophoresis. Results represent the mean of exon 3 skipping level □ SEM of three independent experiments.

[3] The severity (partial or total) of the splicing defects induced by the variation of interest are indicated (total effect, $\Delta 3 \geq 95\%$; partial effect, $95\% \geq \Delta 3 \geq 5\%$).

[4] Complementation by hBRCA2 variants of the ES cell lethal phenotype imposed by Cre-mediated loss of Brca2, visualized by methylene blue staining of HATr colonies of ES cells expressing the different variants of BRCA2 (Figure 2).

[5] Sensitivity of HATr colonies of ES cells expressing hBRCA2 variants to different DNA-damaging agents (Figure 3). n/a, not applicable since no HATr clones were formed after loss of the conditional Brca2 allele.

[6] RT-PCR analyses of patients' RNA are specified when available (Caputo *et al.*, 2018[1]; Théry *et al.*, 2011[2]; Mesman *et al.*, 2013[3]; Biswas *et al.*, 2012[4]; Colombo *et al.*, 2018[5]; Sanz *et al.*, 2010[6]; Thomassen *et al.*, 2012[7]; Bonnet *et al.*, 2008[8] ; Machado *et al.*, 2017[9]).

Figure 1. Comparative RT-PCR analysis of the splicing pattern of *BRCA2* exon 3 in different systems expressing h*BRCA2* SNVs of interest. (A) Analysis of the splicing pattern of pCAS2-*BRCA2*e3 minigenes carrying the SNVs of interest by RT-PCR. For each minigene, the relative quantification of splicing events was evaluated by fluorescent RT-PCR followed by capillary electrophoresis and represent the mean ± SEM of three independent transfection experiment. The identities of the RT-PCR products obtained are indicated on the right. (B) Detection of *BRCA2* exon 3 splicing alterations in patient blood-derived samples. The splicing patterns were monitored by fluorescent RT-PCR followed by capillary electrophoresis and compared to those from at least 3 equivalent control individuals. The relative quantification of splicing events represent the mean ± SEM of three independent experiment. The identities of the RT-PCR products obtained are indicated on the right. (C) Detection of *BRCA2* exon 3 splicing alterations in mouse ES cells expressing hBRCA2 transgene. The splicing patterns of HAT^r mouse ES cells expressing mutant hBRCA2 transgene were monitored by fluorescent RT-PCR followed by capillary electrophoresis and compared to that of the HAT^r mouse ES cells expressing WT hBRCA2 transgene. The relative

quantification of splicing events represent the mean ± SEM of three independent experiment conducted on two clones expressing similar level of hBRCA2 by mutations. The identities of the RT-PCR products obtained are indicated on the right.

Figure 2. Representative images of complementation phenotypes of controls and ES cells expressing the h*BRCA2* SNVs of interest. ES cells expressing WT BRCA2 or BRCA2 variants were electroporated with a pGKCre expression plasmid to induce loss of the conditional BRCA2 allele and restore the HPRT gene. Upon Cre-recombinase expression cells become BRCA2 deficient, which is lethal unless complemented by the expression of a (partially) functional BRCA2 variant. HATr colonies of ES cells expressing no BAC, WT or different variants of BRCA2 were then stained by methylene blue.

Figure 3. Representative graphics of sensitivity of mES cells expressing the h*BRCA2* SNVs of interest to different DNA-damaging agents. Survival of ES-cells expressing c.102A>G or c.316+6T>A mutants (two clones expressing comparable level of BRCA2 by mutations) compared to those expressing wild-type hBRCA2 after exposure to different doses of genotoxins: Olaparib (PARPi, a), Methyl methanesulfonate (MMS, b), Mitomycin C (MMC, c), Cisplatin (d), Camptothesin (e), Ionizing radiation (f). Drug sensitivity was measured by XTT assay and expressed as a percentage of surviving cells compared with untreated cultures. Error bars indicate standard deviation. Differences between the c.316+6T>A mutant and the WT were significant for all DNA-damaging agents whereas differences between c.102A>G and the WT were not significant for all DNA-damaging.

Figure S1. Structure of the pCAS2-*BRCA2*e3 minigene used in the splicing reporter assay. Boxes represent exons, lines in between indicate introns, arrows below the exons represent primers used in RT-PCR reactions, the star designate the 6-FAM fluorochrome used for the capillary electrophoresis, the black bent arrow specifies the cytomegalovirus (CMV) promoter and the black circle indicates the polyadenylation site (Poly A). The pCAS2 vector carries two exons (here named A and B) with a sequence derived from the human *SERPING1/C1NH* gene, separated by an intron containing BamHI and MluI cloning sites. The pCAS2-*BRCA2*e3 minigenes were generated by inserting a genomic fragment containing *BRCA2* exons 3 as well as upstream/downstream flanking intronic sequences (165 and 225 nucleotides, respectively) into the intron of pCAS2 vector, by using BamHI and MluI restriction site. Expression of the pCAS2 minigene is under the

control of a CMV promoter. The pCAS2 is a modified version of the previously described pCAS1 plasmid (Tournier *et al.*, 2008). Two modifications were introduced into the exon A of pCAS2 relative to pCAS1: (i) the first 114 bp of exon A were deleted, and (ii) the SERPING1/CINH translation initiation codon was disrupted by replacing the sequence GATG (initiation codon) by TCAC.

Figure S2. Bioinformatics predictions of variants leading to the use of a de novo or a cryptic splice site. (A) The variant c.68-1G>A induce a total splicing with the production of transcripts containing the *BRCA2* exon 3 deleted of its six first nucleotides (E3Δp(6nt)) due to the creation of a de novo 3' splice site (MES = 8.3) concomitant to the destruction of the natural 3' splice site (MES = 6.1 and 0; SSFL = 87.9 and 0, respectively in the WT and mutated context), as suggested by both MES and SSFL algorithms. (B) The c.100G>A is responsible of a partial splicing defect with the production of three transcripts: (i) one containing the entire exon 3 (E3= 77%), (ii) one without the exon 7 (Δ3= 14%) and (iii) one with inclusion of exon 7 deleted from its first 45 nucleotides ([E3Δp(45nt)] = 9%), due to the creation of a de novo 3' splice site at the position c.112 (MES = 3.2 and SSFL = 77.3) used along with the natural 3' splice site (MES = 6.1 and SSFL = 87.9), as suggested by both MES and SSFL algorithms.

Figure S3. Sanger sequencing of the RT-PCR products of patients carrying variation within *BRCA2* exon 3. Partial sequence chromatograms of the cDNA amplified fragment showing the presence of both allele, wild-type (WT) and mutant (indicated by an arrow) in the RT-PCR products.

Figure S4. Correlation analysis between exon 3 skipping levels observed in pCAS2-BRCA2e3 minigene assay, patient-derived RNA samples or mouse embryonic stem cells. The impact on splicing of *BRCA2* exon 3 variants was determined in the context of the pCAS2-*BRCA2*e3 minigene and compared with patient blood-derived RNA samples (B) or HAT^r mouse embryonic stem (mES) cells (C) when available. The precise correspondence between the levels of exon skipping observed in the pCAS2-*BRCA2*e3 minigene assays, in patient-derived mRNA samples (lymphoblastoid cell lines treated with puromycine and PAXgene) and in HATr mouse embryonic stem cells and the identity of the corresponding *BRCA2* exon 7 variant, is indicated on Table S1. R^2, coefficient of determination (R^2).

Figure S5. Selection of translationally silent variations causing increasing levels of exon 3 skipping according to the results obtained in the pCAS2-BRCA2e3 minigene-based splicing assay. The level of BRCA2 exon 3 skipping induced by translationally silent variants (intronic and synonymous) was determined in the context of the pCAS2-BRCA2e3 minigene. The graph shows the separation of the variants into 3 groups according to the severity of the exon 3 skipping level: (i) variants associated with near to total exon skipping ($\Delta 3 \geq 95\%$, n=9) and reported as pathogenic by Caputo *et al.*, 2018 (red box), (ii) variants associated with weak exon skipping ($\Delta 3 \leq 7\%$, n=29) and reported as neutral by Colombo *et al.*, 2018 (green box) and (iii) variants associated with moderate exon skipping ($7\% \leq \Delta 3 \leq 95\%$, n= 35) and considered as variants of uncertain significance (VUS, grey box). The 9 variants causing increasing levels of exon 3 skipping retrained for protein function analysis are indicated in grey.

Table S1. Description of *BRCA2* exon 3 variants selected in this study. All nucleotide changes identified in *BRCA2* exon 3 and its flanking introns (c.68-8 to c.631+50) after genetic testing of patients reporting strong family history evocative of a HBOC syndrome and reported within the BRCA-Share database by the French BRCA diagnostic laboratories were retrieved (Béroud *et al.*, 2016; Caputo *et al.*, 2012). In addition, all translationally silent single nucleotide variation (intronic and synonymous exonic) mapping *BRCA2* exon 3 and the 10 bp flanking intronic regions (c.68-10 to c.631+10) and potentially affecting s partially plicing were selected based on splicing-dedicated bioinformatics predictions. * Artificial nucleotide variations not yet reported in one of the database of interest.

[1] Variant classification is indicated when attributed and was retrieved from each database and refers to the 5-tier system used by the InSiGHT Variant Interpretation Committee (http://insight-group.org/variants/classifications/) as follows: 1, not pathogenic or begnign; 2, likely not pathogenic or likely begnign; 3, uncertain significance (also called VUS for variants of unknown significance); 4, likely pathogenic; 5, pathogenic; n/a, not available.

[2] For each splicing functional assay, the relative quantification of splicing events was evaluated by fluorescent RT-PCR followed by capillary electrophoresis. Results represent the mean of exon 3 skipping level ± SEM of three independent experiments.

[3] Variants producing exon skipping levels greater than the one of the WT (1 ± 5%) were considered as splicing mutations (increased exon skipping). The severity (partial or total) of the splicing defects induced by the variation of interest are indicated (total effect, Δ3 ≥ 95%; partial effect, 95% ≥ Δ3 ≥ 5%), as well as the nature of the splicing defects (Δ3, increased exon skipping; E3Δp(1nt), deletion of the first nucleotide; E3Δp(45nt), deletion of the first 45 nt). When the variation of interest induce several splicing defects, the identities and proportion of the abnormal RT-PCR products are specified.

[4] Complementation by hBRCA2 variants of the ES cell lethal phenotype imposed by Cre-mediated loss of Brca2, visualized by methylene blue staining of HATr colonies of ES cells expressing the different variants of BRCA2 (Figure 2).

[5] Sensitivity of HATr colonies of ES cells expressing hBRCA2 variants to different DNA-damaging agents (Figure 3). n/a, not applicable since no HATr clones were formed after loss of the conditional Brca2 allele.

[6] RT-PCR analyses of patients' RNA are specified when available (Caputo *et al.*, 2018[1]; Théry *et al.*, 2011[2]; Mesman *et al.*, 2013[3]; Biswas *et al.*, 2012[4]; Colombo *et al.*, 2018[5]; Sanz *et al.*, 2010[6]; Thomassen *et al.*, 2012[7]; Bonnet *et al.*, 2008[8] ; Machado *et al.*, 2017[9]).

Table S2. Description of the primers used in this study.

[1] F, forward; R, reverse.

[2] Intronic and exonic sequences are indicated in grey and black, respectively. The position of the nucleotide variation is underlined. The double underlined sequences correspond to restriction sites for BamHI and MluI and the sequence highlighted in grey correspond to the 15bp-tail used for homologous recombination.

Table S3. Comparison of experimental results obtained with pCAS2-BRCA2e3 minigenes carrying natural variants mapping to *BRCA2* exon 3 splice sites or to flanking intronic positions with *in silico* data obtained with splice site-dedicated bioinformatics approaches. The impact on splicing of 27 variants located in or near the splice sites of *BRCA2* exon 3 was

determined by performing a cell-based splicing assay with pCAS2-*BRCA2*e7 minigenes or patient RNA analysis, when available, as shown in Table 1. Δ3p(6nt), inclusion of exon 3 deleted of its sox first nucleotides. The table shows a separation of the variants into 2 groups according to the minigene results shown in Figure 1: variants that induced splicing defects (n=23) and those that did not (n=4). *In silico* predictions of potential effects on splicing were conducted by using 2 splice site-dedicated *in silico* tools (MES, SSFL), as well as two approaches based on their combination (MES+SSFL and SPiCE). MES and SSFL results are presented as the change in scores (Δ) of the variants relative to WT (ΔMES and ΔSSFL, respectively). True and false calls (in grey) of exon 3 splicing defects were determined by taking into account the following thresholds: -15% for ΔMES, -5% for ΔSSFL and 11.5% for SPiCE as previously recommended (Houdayer *et al.*, 2012; Leman *et al.*, 2018) and indicated between parenthesis in the Table.

Table S4. Comparison of the experimental results obtained with pCAS2-*BRCA2*e3 minigenes carrying natural *BRCA2* exon 3 variants with *in silico* data obtained with SRE-dedicated bioinformatics approaches. The impact on splicing of 70 *BRCA2* exon 3 variants was determined by performing a cell-based splicing assay with pCAS2-BRCA2e3 minigenes as shown in Table 1. The table shows a separation of the variants into 2 groups according to the minigene results shown in Table 1: variants that induced exon 3 skipping (n=23) and variations with no effect on exon 3 splicing (n = 14). *In silico* predictions of potential effects on ESRs were conducted by using the 4 new SRE-dedicated *in silico* tools (QUEPASA, HEXplorer, SPANR and HAL), as well as three approaches resulting from their combination (QUEPASA&HAL, at AT LEAST 3 and LR_{skip}). True and false calls (in grey) for prediction of induced exon-skipping events were determined by taking into account the following thresholds as previously recommended (Tubeuf *et al.*, in preparation) and indicated between parenthesis in the Table: -0.50 for QUEPASA (ΔtESRseq scores), -14 for HEXplorer (ΔHZEI scores), -0.1% for SPANR (Δψ scores), -3.4% for HAL (Δψ scores) and 31.1% for LR_{skip}. n/a, not applicable.

Variations		Databases Clinical classification[1]									RNA splicing assay[2]						Protein functional assay			
Nucleotide variations	Predicted protein changes	COSMIC	ClinVar	dbSNP	ESP	gnmoAD	HGMD	BIC	LOVD	BRCA-Share	Minigene	Patient RNA analysis			HATr mESC	Effect on splicing[3]	Complementation[4]	Sensitivity to DNA damage agents[5]	Classification suggested	References[6]
												LCL (-/+ puro)	PAXgene	ASE						
WT											99 ± 0.2	92 ± 2.3 88 ± 1.9	89 ± 1.1	100	90 ± 0.3		Yes	No		
c.68-7T>A	p.?	n/a	1 2 3	1 2	n/a	n/a		3	n/a	3	92 ± 1.1	[75 ± 0.2 67 ± 0.2] [75 ± 0.2 72 ± 0.4] [75 ± 0.3 65 ± 0.3] [74 ± 0.4 67 ± 0.5]				Partial effect (Δ3)	Yes	No	1	[2,5]
c.68-7_68-8delinsAA	p.?		3							2	55 ± 2.2		47 ± 0.9		45 ± 1.1	Partial effect (Δ3)	Yes	No	1	
c.102A>G	p.=		2	2		n/a				3	51 ± 1.4		45 ± 1.3	54 ± 1.9	44 ± 0.9	Partial effect (Δ3)	Yes	No	1	
c.165C>T	p.=		2	4							85 ± 1.6				69 ± 0.2	Partial effect (Δ3)	Yes	No	1	
c. 231T>G	p.=		1 2	1	n/a	n/a			n/a	3	69 ± 3.4	[62 ± 4.8 60 ± 5.2][a] [56 ± 4.8 54 ± 5.2][b] [55 ± 4.8 53 ± 5.2][c] [48 ± 4.8 44 ± 5.2][d]		65 ± 1.1[a] 62 ± 1.3[b] 56 ± 0.9[c] 108 ± 2.1[d]	61 ± 1.1	Partial effect (Δ3)	Yes	No	1	[2]
c.316G>A	p.Gly106Arg		3	3		n/a				5	58 ± 1.2	[39 ± 0.5 39 ± 0.5][e] [39 ± 0.5 36 ± 0.7][f]		22 ± 0.9[e] 24 ± 0.5[f]		Partial effect (Δ3)			3	
c.316+5G>A	p.?		4 5	4 5						5	5 ± 0.3					Total effect (Δ3)			5	[1,7]
c.316+6T>A*	p.?										26 ± 0.4				15 ± 1.4	Partial effect (Δ3)	Poor	Yes	5	
c.316+6T>C	p.?		3	3		n/a		3		5	35 ± 2.3	38 ± 2.1 40 ± 1.9			24 ± 0.2	Partial effect (Δ3)	Yes	No	1	
c.316+6T>G	p.?					n/a					23 ± 2.1				11 ± 1.1	Partial effect (Δ3)	Poor	Yes	5	

Table 1. Functional evaluation of *BRCA2* exon 3 variants causing increasing levels of exon 3 skipping.

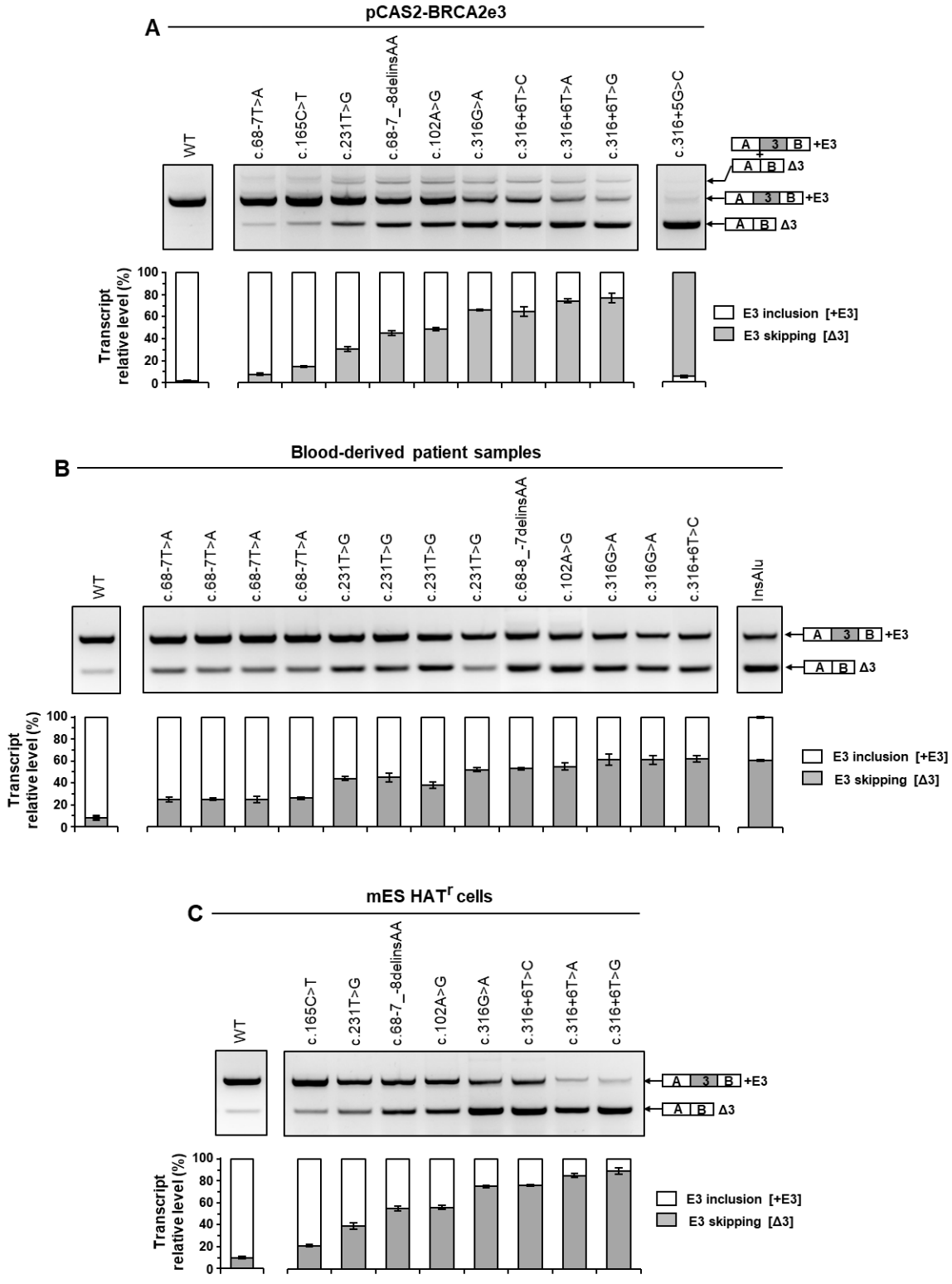


Figure 1. Comparative RT-PCR analysis of the splicing pattern of *BRCA2* exon 3 in different systems expressing the h*BRCA2* SNVs of interest.

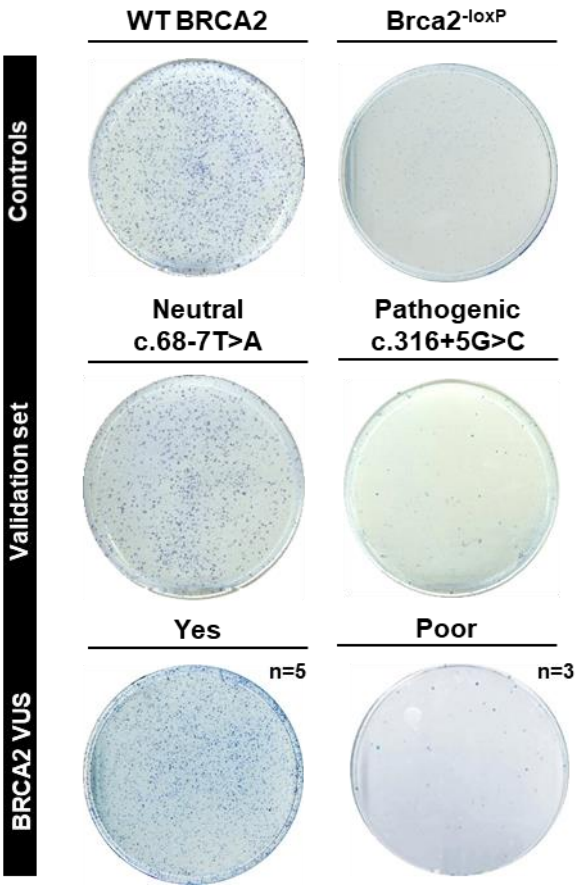


Figure 2. Representative images of complementation phenotypes of controls and ES cells expressing the h*BRCA2* SNVs of interest.

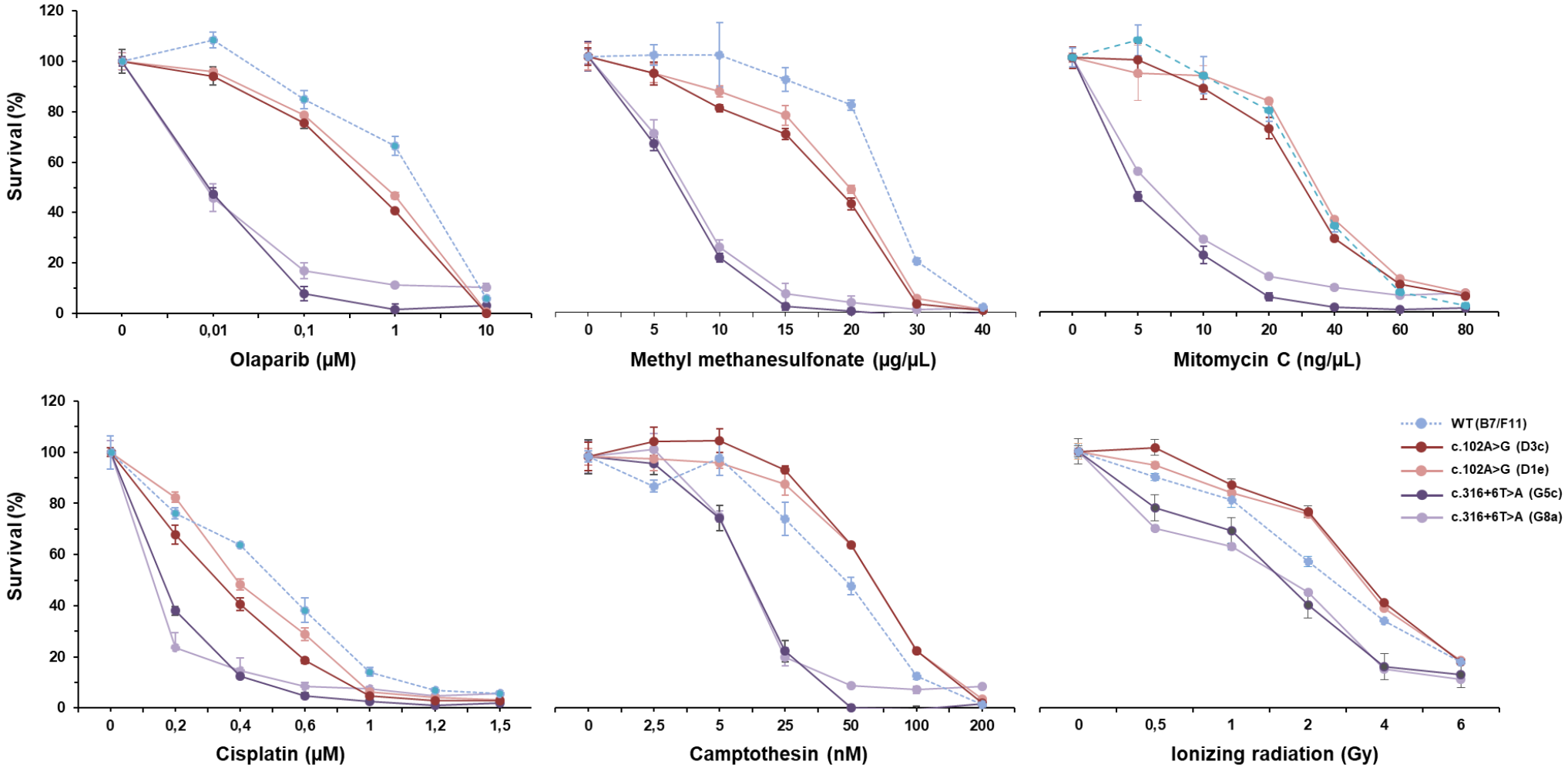


Figure 3. Representative graphics of sensitivity of mES cells expressing translationally silent h*BRCA2* SNVs of interest to different DNA-damaging.

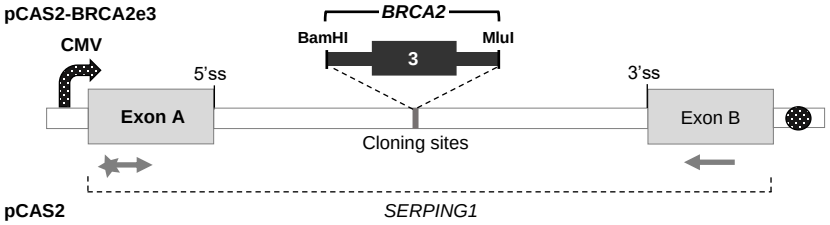


Figure S1. Structure of the pCAS2-*BRCA2*e3 minigene used in the splicing reporter assay.

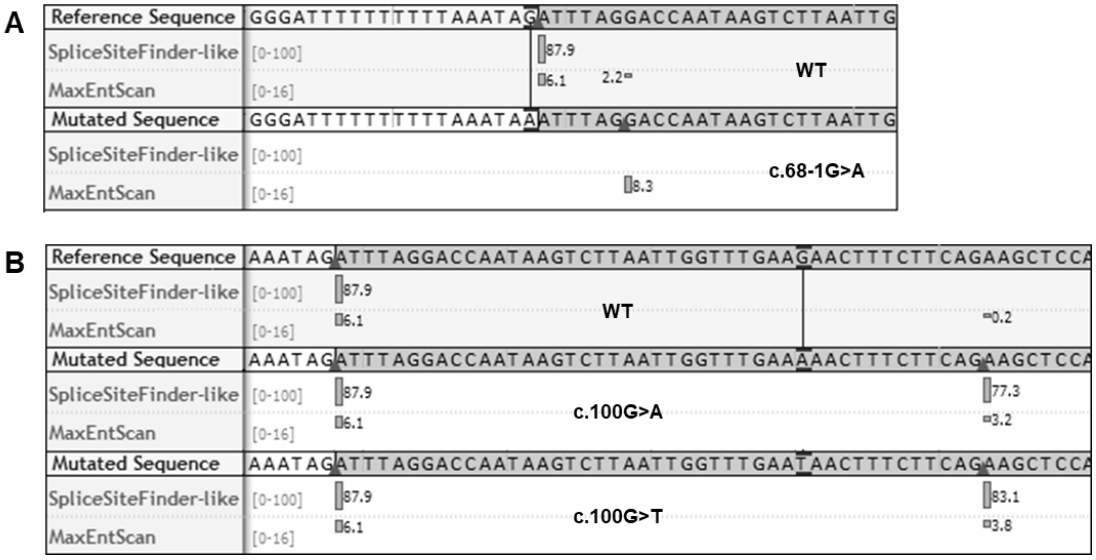


Figure S2. Bioinformatics predictions of variants leading to the use of a de novo or a cryptic splice site.

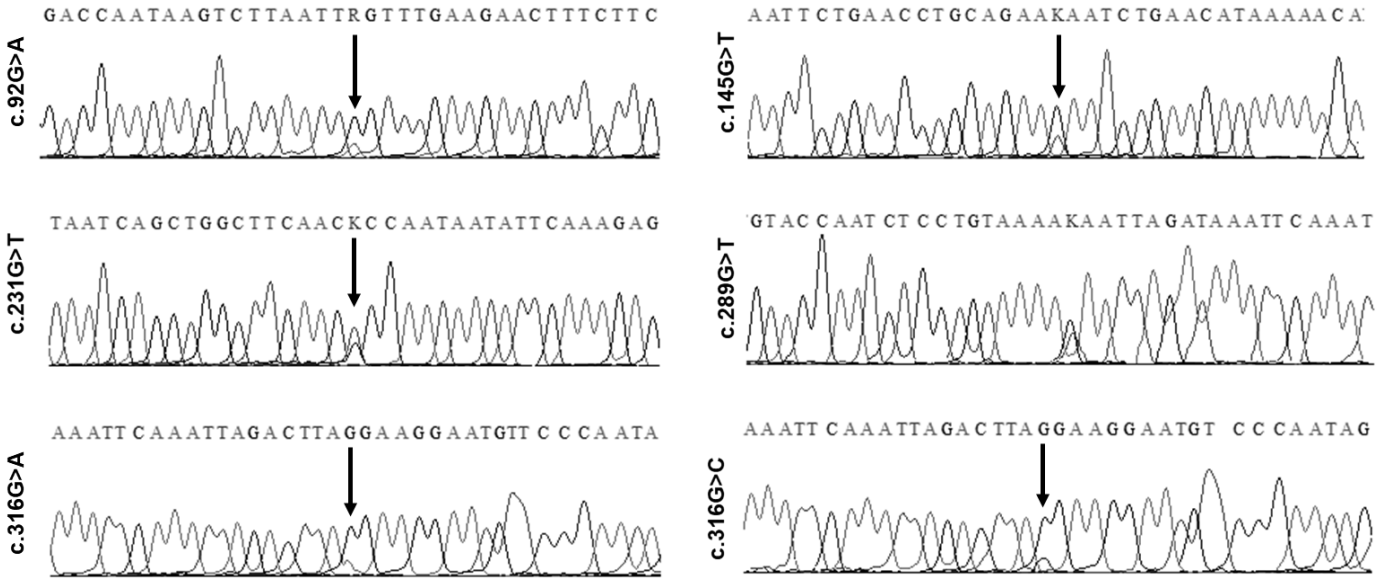


Figure S3. Sanger sequencing of the RT-PCR products of patients carrying variation within *BRCA2* exon 3.

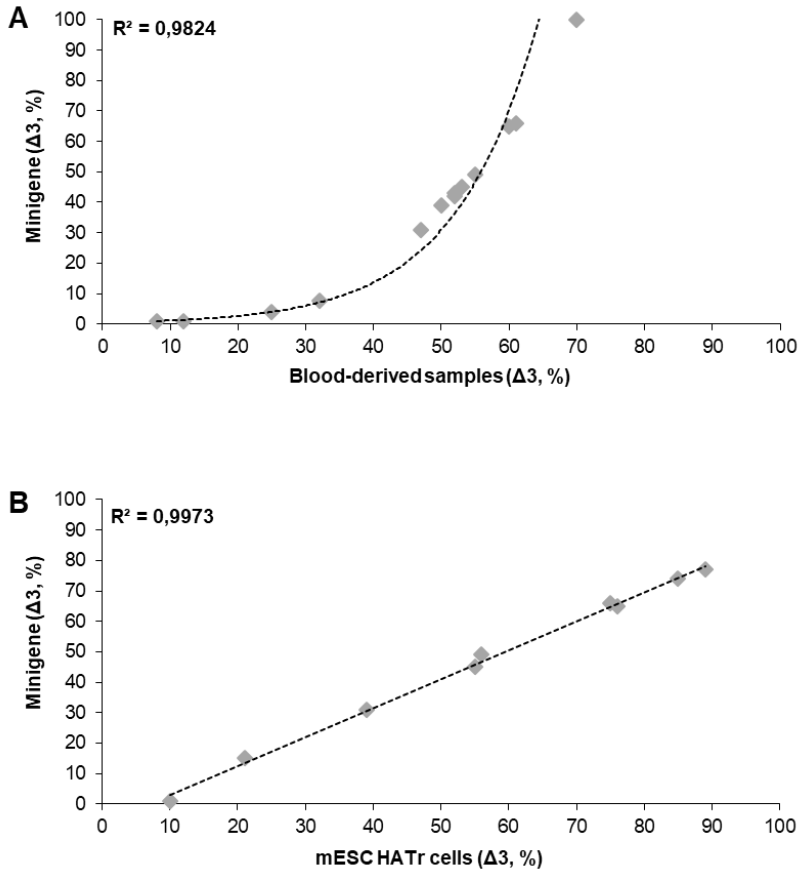


Figure S4. Correlation analysis between exon 3 skipping levels observed in pCAS2-*BRCA2*e3 minigene assay, patient-derived RNA samples or mouse embryonic stem cells.

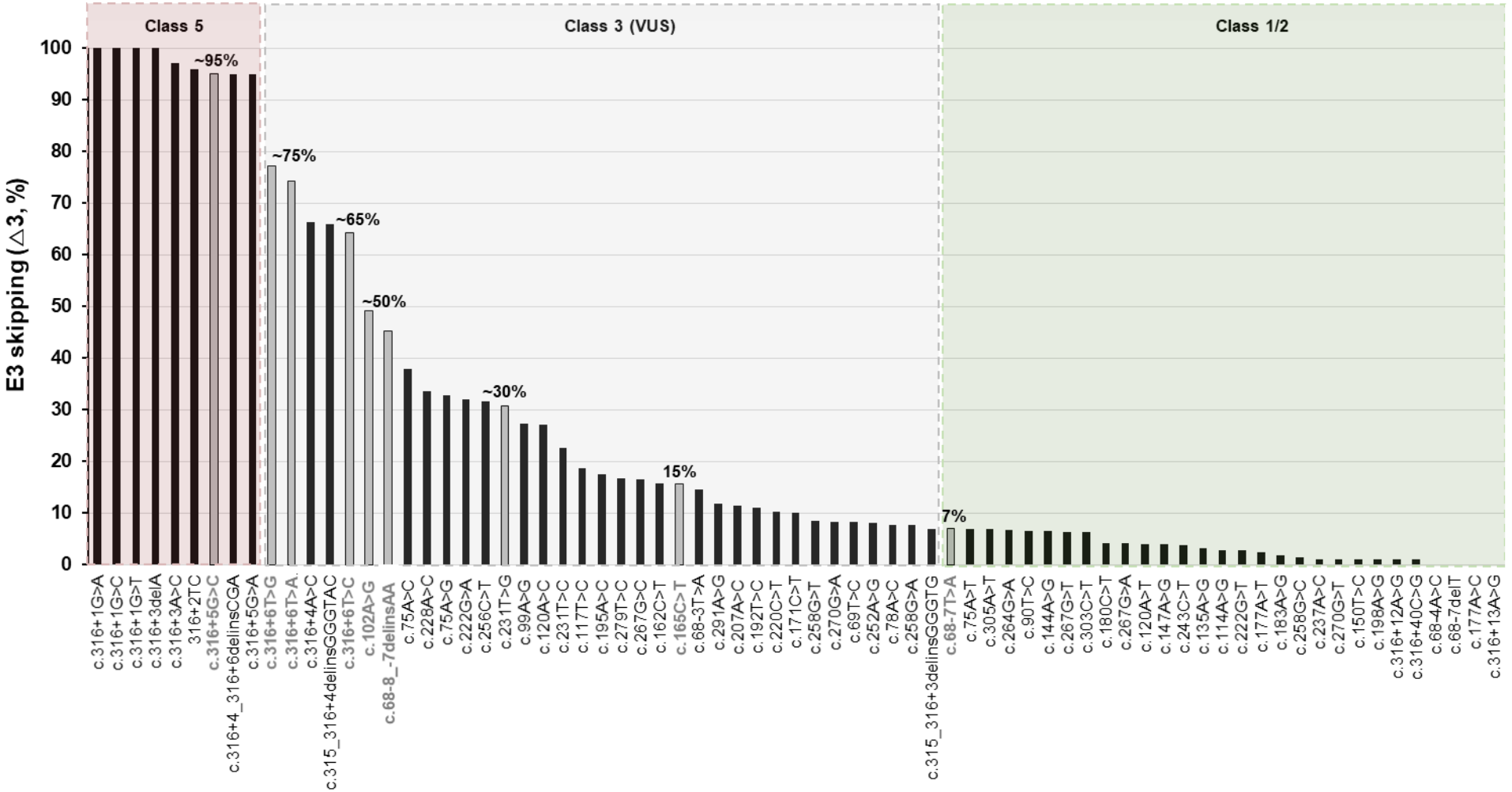


Figure S5. Selection of translationally silent variations causing increasing levels of exon 3 skipping according to the results obtained in the pCAS2-BRCA2e3 minigene-based splicing assay.

Variations			Databases Clinical classification[1]									RNA splicing assay[2]						Protein functional assay			
Positions	Nucleotide variations	Predicted protein changes	COSMIC	ClinVar	dbSNP	ESP	gnmoAD	HGMD	BIC	LOVD	BRCA-Share	Minigene	Patient RNA analysis			HATr mESC	Effect on splicing[3]	Complementation[4]	Sensitivity to DNA damage agents[5]	Classification suggested	References[6]
													LCL (-/+ puro)	PÄXgene	ASE						
	WT											99 ± 0.2	92 ± 2.3 88 ± 1.9	89 ± 1.1	100	90 ± 0.3		Yes	No		
	c.157insalu	p.Asp23_Leu105del								n/a	n/a	nd	30 ± 0.1 30 ± 0.1				Total effect (Δ3)			5	[1,9]
Intron 2 (n=6)	c.68-7delT	-	n/a	1 3	n/a	n/a	n/a		3	n/a	1	99 ± 0.3					No effect			1	[2]
	c.68-7T>A	p.?	n/a	1 2 3	1 2	n/a	n/a		3	n/a	3	92 ± 1.1	[75 ± 0.2 67 ± 0.2] [75 ± 0.2 72 ± 0.4] [75 ± 0.3 65 ± 0.3] [74 ± 0.4 67 ± 0.5]				Partial effect (Δ3)	Yes	No	1	[2,5]
	c.68-7_68-8delinsAA	p.?		3							2	55 ± 2.2		47 ± 0.9		45 ± 1.1	Partial effect (Δ3)	Yes	No	1	
	c.68-4A>C*	p.?										100 ± 0.1					No effect			1	
	c.68-3T>A*	p.?										85 ± 0.6					Partial effect (Δ3)			1	
	c.68-1G>A*	p.?		4								99 ± 1.0					Total effect (E3Δp(6nt))			3	
Exon 3 (n=82)	c.69T>C	p.=		2						n/a		92 ± 0.7					Partial effect (Δ3)			1	
	c.74G>T	p.Gly25Val									3	88 ± 0.1					Partial effect (Δ3)			3	
	c.75A>G	p.=	n/a									67 ± 0.8					Partial effect (Δ3)			1	
	c.75A>C*	p.=										63 ± 3.4					Partial effect (Δ3)			1	
	c.75A>T*	p.=										93 ± 0.8					Partial effect (Δ3)			1	
	c.78A>C*	p.=										92 ± 0.9					Partial effect (Δ3)			1	
	c.79A>G	p.Ile27Val		3	3		n/a	4	3	n/a	3	94 ± 0.2					Partial effect (Δ3)			3	
	c.81A>T*	p.=										99 ± 0.1					No effect			3	
	c.90T>C	p.=		2 3	2			n/a				95 ± 0.3					Partial effect (Δ3)			1	
	c.92G>C	p.Thr31Ser									5	99 ± 0.0					No effect			3	
	c.92G>A	p.Trp31*		5	3 5		n/a	5			5	57 ± 0.2	59 ± 1.9 50 ± 1.7		48 ± 1.1		Partial effect (Δ3)			5	
	c.99A>G	p.=	n/a									73 ± 1.7					Partial effect (Δ3)			1	
	c.100G>A	p.Glu34Lys					n/a				3	86 ± 0.3					Partial effect (Δ3 = 14% ; c = 9%)			3	

	c.102A>G	p.=		2	2		n/a				3	51 ± 1.4		45 ± 1.3	54 ± 1.9	44 ± 0.9	Partial effect (Δ3)	Yes	No	1	
	c.107C>T	p.Ser36Phe		3	3		n/a				3	79 ± 0.3					Partial effect (Δ3)			3	
	c.114A>G	p.=		2							2	97 ± 0.2					No effect			1	
	c.117T>C*	p.=										92 ± 2.8					Partial effect (Δ3)			1	
	c.120A>C*	p.=										73 ± 3.1					Partial effect (Δ3)			1	
	c.120A>T*	p.=										96 ± 0.5					No effect			1	
	c.121C>T	p.Pro41Ser		3	3		n/a		3		3	99 ± 0.3					No effect			3	
	c.122C>T	p.Pro41Ser	n/a	1 2 3	1 3		n/a	5			3	99 ± 0.1					No effect			3	
	c.135A>G*	p.=										97 ± 0.4					No effect			1	
	c.139G>T	p.Ala47Ser									3	91 ± 0.7					Partial effect (Δ3)			3	
	c.140C>T	p.Ala47Val									3	92 ± 0.3					Partial effect (Δ3)			3	
	c.142G>A	p.Gly48Lys		3							3	92 ± 0.4					Partial effect (Δ3)			3	
	c.143A>G	p.Glu48Gly									3	93 ± 0.2					Partial effect (Δ3)			3	
	c.144A>G	p.=										93 ± 0.3					Partial effect (Δ3)			1	
	c.145G>T	p.Glu49*		5	4 5		n/a	5	5	n/a	5	62 ± 0.1	61 ± 1.1 53 ± 0.9		50 ± 1.2		Partial effect (Δ3)			5	[6]
	c.147A>G*	p.=										94 ± 0.2					No effect			1	
	c.150T>C	p.=									3	99 ± 0.1					No effect			1	
	c.162C>T*	p.=										84 ± 1.2					Partial effect (Δ3)			1	
	c.165C>T	p.=		2	4							85 ± 1.6				79 ± 0.2	Partial effect (Δ3)	Yes	No	1	
	c.167A>C	p.Asn56Thr		1 2	1 2		n/a		3	n/a	1	99 ± 0.3					No effect			3	
	c.171C>T	p.=		2 3	2 3 5		n/a				3	92 ± 0.8					Partial effect (Δ3)			1	
	c.175C>G	p.pro59ala		1 2 3	1 3	n/a	n/a		3		3	100 ± 0.0					No effect			3	
	c.177A>C	p.=									3	100 ± 0.1					No effect			1	
	c.177A>T*	p.=										98 ± 0.2					No effect			1	
	c.179A>G	p.Asn60Ser		2 3	2		n/a		3		3	99 ± 0.2					No effect			3	
	c.180C>T	p.=		2			n/a					96 ± 0.3					No effect			1	
	c.183A>G	p.=		2 3			n/a				3	98 ± 0.5					No effect			1	
	c.191C>T	p.Thr64Ile		3	3		n/a	4			3	99 ± 0.5					No effect			3	
	c.192T>C*	p.=										89 ± 1.4					Partial effect (Δ3)			1	
	c.195A>C	p.=		2								82 ± 1.7					Partial effect (Δ3)			1	
	c.198A>G	p.=	n/a	1 2	1 2	n/a	n/a	4	3		2	99 ± 0.1					No effect			1	
	c.207A>C*	p.=										88 ± 1.1					Partial effect (Δ3)			1	
	c.220C>T*	p.=										90 ± 0.2					Partial effect (Δ3)			1	
	c.222G>A	p.=		2			n/a					68 ± 2.1					Partial effect (Δ3)			1	
	c.222G>T*	p.=										97 ± 0.1					No effect			1	
	c. 223G>C	p.Ala75Pro		1 2 3	1 2	n/a	n/a	4	3	n/a	1	99 ± 0.1	91 ± 0.2 87 ± 0.6		96 ± 2.1		No effect			3	[2]
	c.228A>C*	p.=										66 ± 0.4					Partial effect (Δ3)			1	
	c. 231T>G	p.=		1 2	1	n/a	n/a			n/a	3	69 ± 3.4	[62 ± 4.8		65 ± 1.1[a]	61 ± 1.1	Partial effect (Δ3)	Yes	No	1	[2]

														60 ± 5.2][a] [56 ± 4.8 54 ± 5.2][b] [55 ± 4.8 53 ± 5.2][c] [48 ± 4.8 44 ± 5.2][d]		62 ± 1.3[b] 56 ± 0.9[c] 108 ± 2.1[d]						
	c.231T>C	p.=									3	77 ± 0.7					Partial effect (Δ3)			1	
	c.237A>C	p.=									3	99 ± 0.2					No effect			1	
	c.240A>G	p.Ile80Met		3		3				3		3	99 ± 0.3					No effect			3	
	c.241T>A	p.Phe81Ile		2	3	2	3		n/a	4	3		3	99 ± 0.3					No effect			3	
	c.243C>T	p.=		2										96 ± 0.3					No effect			1	
	c.252A>G*	p.=												92 ± 0.9					Partial effect (Δ3)			1	
	c.256C>T	p.=		2										68 ± 1.2					Partial effect (Δ3)			1	
	c.258G>A*	p.=												92 ± 0.3					Partial effect (Δ3)			1	
	c.258G>T*	p.=												91 ± 0.4					Partial effect (Δ3)			1	
	c.258G>C*	p.=												98 ± 0.0					No effect			1	
	c.260C>G	p.Thr87Ser											3	97 ± 0.2					No effect			3	
	c.264G>A*	p.=												92 ± 0.1					Partial effect (Δ3)			1	
	c.266C>T	p.Pro89Leu	n/a	3		n/a			n/a				3	97 ± 0.3					No effect			3	
	c.267G>C*	p.=												83 ± 1.2					Partial effect (Δ3)			1	
	c.267G>T	p.=		2		1	2		n/a					94 ± 0.7					Partial effect (Δ3)			1	
	c.267G>A	p.=		1	2	3	1	2		n/a				3	96 ± 0.2					No effect			1	
	c.270G>T*	p.=												99 ± 0					No effect			1	
	c.270G>A*	p.=												92 ± 0.7					Partial effect (Δ3)			1	
	c.279T>C*	p.=												83 ± 0.5					Partial effect (Δ3)			1	
	c.280C>T	p.Pro94Ser		2	3	2	3		n/a		3		3	97 ± 0.7					No effect			3	
	c.289G>T	p.Glu97*	n/a	5		5							5	nd	77 ± 1.3 75 ± 1.3		52 ± 0.6		Partial effect (Δ3)			5	
	c.291A>G*	p.=												88 ± 2.1					Partial effect (Δ3)			1	
	c.303C>T	p.=		2		3								94 ± 0.2					Partial effect (Δ3)			1	
	c.305A>T	p.Lys102Ile											3	98 ± 0.3					No effect			3	
	c.315_316+3delinsGGGTG*	p.Gly106dup												93 ± 0.9					Partial effect (Δ3)			3	
	c.315_316+4delinsGGGTAC*	p.Arg107Thrfs*7												37 ± 3.1					Partial effect (Δ3)			5	
	c.316G>C	p.Gly106Arg											3	58 ± 1.2	56 ± 1.1 48 ± 1.5		39 ± 1.4		Partial effect (Δ3)			3	
	c.316G>A	p.Gly106Arg		3		3			n/a				5	34 ± 0.2	[39 ± 0.5 39 ± 0.5][e] [39 ± 0.5 36 ± 0.7][f]		22 ± 0.9[e] 24 ± 0.5[f]	25 ± 0.6	Partial effect (Δ3)	Poor	Yes	5	
Intron 3 (n=16)	c.316+1G>A	p.?		5		4	5							0 ± 0.0					Total effect (Δ3)			5	
	c.316+1G>C	p.?		5		4	5							0 ± 0.0					Total effect (Δ3)			5	

	c.316+1G>T	p.?	n/a	4	5	4	5							0 ± 0.0					Total effect (Δ3)			5	[1]
	c.316+2T>C	p.?		5		5					3		5	4 ± 0.2					Total effect (Δ3)			5	[1]
	c.316+3A>C	p.?											3	3 ± 0.9					Total effect (Δ3)			5	
	c.316+4delA	p.?		n/a									5	0 ± 0.0					Total effect (Δ3)			5	[1]
	c.316+4A>C*	p.?												34 ± 3.2					Partial effect (Δ3)			1	
	c.316+4_316+6delinsCGA*	p.?												5 ± 1.2					Total effect (Δ3)			5	
	c.316+5G>A	p.?		4	5	4	5						5	5 ± 0.3					Total effect (Δ3)			5	[1,7]
	c.316+5G>C	p.?		5		4	5						5	5 ± 1.6					Total effect (Δ3)	No	n/a	5	[1,8]
	c.316+6T>A*	p.?												26 ± 0.4				15 ± 1.4	Partial effect (Δ3)	Poor	Yes	5	
	c.316+6T>C	p.?		3		3			n/a		3		5	35 ± 2.3	38 ± 2.1 40 ± 1.9			24 ± 0.2	Partial effect (Δ3)	Yes	No	1	
	c.316+6T>G	p.?							n/a					23 ± 2.1				11 ± 1.1	Partial effect (Δ3)	Poor	Yes	5	
	c.316+12A>G	p.?		1	2	2			n/a				3	99 ± 0.3					No effect			1	
	c. 316+13A>G	p.?		1	2	3	3		n/a				3	100 ± 0.1					No effect			1	
	c.316+40C>G	p.?											3	99 ± 0.0					No effect			1	

Table S1. Description of *BRCA2* exon 3 variants selected in this study.

Purpose		Name[1]	Sequence[2]
Splicing reporter minigene-based assay	Site-directed mutagenesis	Br2Ex3_c.68-4AC-R	CCTAAATCTAGTTAAAAAAAAAATC
		Br2Ex3_c.68-3TA-F	TTTTTTTTTAAAAAGATTTAGGACC
		Br2Ex3_c.68-1GA-F	TTTTAAATAAATTTAGGACCAATAAG
		Br2Ex3_c.69TC-F	TTTTAAATAGACTTAGGACCAATAAG
		Br2Ex3_73GA-F	AATAGATTTAAGACCAATAAGTC
		Br2Ex3_75AC-F	AATAGATTTAGGCCCAATAAGTC
		Br2Ex3_75AG-R	GACTTATTGGCCCTAAATCTATT
		Br2Ex3_75AT-R	GACTTATTGGACCTAAATCTATT
		Br2Ex3_78AC-F	GATTTAGGACCCATAAGTCTTAA
		Br2Ex3_81AT-F	AGGACCAATTAGTCTTAATTGG
		Br2Ex3_91TC-R	GTTCTTCAAACCGATTAAGACTTATTG
		Br2Ex3_93GT-R	GAAAGTTCTTCAAAACAATTAAGAC
		Br2Ex3_99AG-F	AATTGGTTTGAGGAACTTTCTTC
		Br2Ex3_100GT-R	CTGAAGAAAGTTATTCAAACCAATT
		Br2Ex3_107CT-R	AGCTTCTGAAAAAAGTTCTTC
		Br2Ex3_114AG-F	CTTTCTTCAGAGGCTCCACCCTA
		Br2Ex3_117TC-R	ATTATAGGGTGGGGCTTCTGAAG
		Br2Ex3_120AC-F	TCAGAAGCTCCCCCCTATAATTC
		Br2Ex3_120AT-R	GAATTATAGGGAGGAGCTTCTGA
		Br2Ex3_135AG-F	CTATAATTCTGAGCCTGCAGAAG
		Br2Ex3_144AG-R	TGTTCAGATTCCTCTGCAGGTTC
		Br2Ex3_147AG-F	CCTGCAGAAGAGTCTGAACATA
		Br2Ex3_162CT-R	TCGTAATTGTTATTTTTATGTTC
		Br2Ex3_165CT-F	CATAAAAACAATAATTACGAACC
		Br2Ex3_171CT-F	AACAACAATTATGAACCAAACCT
		Br2Ex3_177AT-R	AATTACGAACCCAACCTATTTAAAAC
		Br2Ex3_180CT-R	TAAATAGGTTAGGTTCGTAATTG
		Br2Ex3_192TC-F	ATTTAAAACCCCACAAAGGAAAC
		Br2Ex3_195AC-R	GATGGTTTCCTTTGGGGAGTTTTA
		Br2Ex3_207AC-F	CAAAGGAAACCCTCTTATAATCAG
		Br2Ex3_220CT-R	GAGTTGAAGCCAACTGATTATAAG
		Br2Ex3_222GA-F	CTTATAATCAGCTAGCTTCAACTC
		Br2Ex3_222GT-R	GAGTTGAAGCAAGCTGATTATAAG
		Br2Ex3_228AC-R	ATTATTGGAGTGGAAGCCAGCTG
		Br2Ex3_231TC-F	GGCTTCAACCCCAATAATATTC
		Br2Ex3_243CT-R	CCTTGCTCTTTAAATATTATTGG
		Br2Ex3_252AG-F	TCAAAGAGCAGGGGCTGACTCTG
		Br2Ex3_256CT-R	GCAGAGTCAACCCTTGCTCTTTG
		Br2Ex3_258GA-F	GAGCAAGGGCTAACTCTGCCGCTGTAC
		Br2Ex3_258GC-F	GAGCAAGGGCTCACTCTGCCGCTGTAC
		Br2Ex3_258GT-R	GTACAGCGGCAGAGTAAGCCCTTGCTC
		Br2Ex3_267GA-F	CTGACTCTGCCAACTGTACCAATC
		Br2Ex3_267GC-F	CTGACTCTGCCCCTGTACCAATC
		Br2Ex3_267GT-R	GATTGGTACAGAGGCAGAGTCAG
		Br2Ex3_270GA-F	ACTCTGCCGCTATACCAATCTCC
		Br2Ex3_270GT-R	GAGATTGGTAAAGCGGCAGAGTC
		Br2Ex3_279TC-F	GCTGTACCAATCCCCTGTAAAAG
		Br2Ex3_291AG-R	GAATTTATCTAACTCTTTTACAGG
		Br2Ex3_303CT-R	CTAAGTCTAATTTAAATTTATCT
		Br2Ex3_315_316+3delinsGGGTG-F	CAAATTAGACTTGGGTGAGTAATGCA

		Br2Ex3_315_316+4delinsGGGTAC	GCATTACGTACCCAAGTCTAATTTG
		Br2Ex3_316+1GA-F	CAAATTAGACTTAGATAAGTAATGC
		Br2Ex3_316+1GC	GCATTACTTAGCTAAGTCTAATTTG
		Br2Ex3_316+1GT-F	CAAATTAGACTTAGTTAAGTAATGC
		Br2Ex3_316+3AC	GCATTACTGACCTAAGTCTAATTTG
		Br2Ex3_316+4AC-F	CAAATTAGACTTAGGTACGTAATGC
		Br2Ex3_316+4_316+6delinsCGA-R	GCATTTCGTACCTAAGTCTAATTTG
		Br2Ex3_316+6TC-R	CTACCATATTGCATTGCTTACCTAAGTC
		Br2Ex3_316+6TA-F	GACTTAGGTAAGAAATGCAATATGGTAG
	PCR (cloning, minigene preparation)	Br2Ex3_InFus_BamHI-F	AGGCTAAGAAGTGCAGGATCTGTTATACCTTTGCCCTGAGATTTAC
		Br2Ex3_InFus_MluI-R	AGGGGTCAAAACAAGACGCGTTGCTTGACACCACTGGACTA
	Sequencing of minigene inserts	pCAS-Seq-F	GGGGTCAATAGCAGTGAGAG
		pCAS-Seq-R	GCTCCATTTCACAGGTAGAGA
	RT-PCR and sequencing of RT-PCR products	6FAM-pCAS-KO1-F (5'-fluo)	TGACGTCGCCGCCCATCAC
		pCAS-2R	ATTGGTTGTTGAGTTGGTTGTC
Allele specific expression (SNaPshot®)	Primer extension	Br2Ex3_c.92GA-F	GAAATATGTTTTGGTGTCTGACGAC
		Br2Ex3_c.92GA-R	TTTCAGAAATATGTTTTGGTGTCTGAC
		Br2Ex3_c.102AG-F	GGACCAATAAGTCTTAATTGGTTTGAAGA
		Br2Ex3_c.102AG-R	TATAGGGTGGAGCTTCTGAAGAAAG
		Br2Ex3_c.145GT-F	CTTTCAGAAATATGTTTTGGTGTCTGA
		Br2Ex3_c.145GT-R	GGTCGTCAGACACCAAAACATAT
		Br2Ex3_c.223GC-F	CCTCAGCTCCTAGACTTTCAGAAA
		Br2Ex3_c.223GC-R	GAAAGTCTAGGAGCTGAGGTGG
		Br2Ex3_c.231TG-F	GAACTTGACCAAGACATATCAGGA
		Br2Ex3_c.231TG-R	GGAGCTGAGGTGGATCCTGA
		Br2Ex3_c.289GT-F	GTAGCTAAAGAACTTGACCAAGACAT
		Br2Ex3_c.289GT-R	GGAGCTGAGGTGGATCCTGATA
		Br2Ex3_c.316G-F	TGTAGCTAAAGAACTTGACCAAGAC
	RT-PCR on patient RNA	Br2Ex2_F	ATGCCTATTGGATCCAAAGAGAGGCC
		Br2Ex5_R	CTGACTTATCTCTTTGTGGTGTTACATG
	PCR on genomic DNA	Br2Ex7_InFus_BamHI-F	AGGCTAAGAAGTGCAGGATCTTTCGCAAGAGAATGGATTAATGATC
		Br2Ex7_InFus_MluI-R	AGGGGTCAAAACAAGACGCGTGGAGGGATGAAAGAGAACATTTAC
Patient RNA analysis	RT-PCR on blood-derived patient RNA	Br2Ex2_F	ATGCCTATTGGATCCAAAGAGAGGCC
		Br2Ex5_R	CTGACTTATCTCTTTGTGGTGTTACATG
Mouse embryonnic stelm cell assay	RT-PCR on RNA to check for *BRCA2* expression in G418^r cells	B2ex11FRT	CCAAGTCATGCCACACATTC
		B2ex14RRT	ATTCTTGACCAGGTGCGGTA
	RT-PCR on RNA to quantify splicing defects in HAT^r cells	Br2Ex2_F	ATGCCTATTGGATCCAAAGAGAGGCC
		Br2Ex5_R	CTGACTTATCTCTTTGTGGTGTTACATG

Table S2. Description of the primers used in this study.

Variations *BRCA2* Exon 3 (n = 27)	Effect on splicing	Exon 3 inclusion (%)	ΔMES (-15%)	ΔSSFL (-5%)	ΔMES and ΔSSFL	SPiCE (11,5%)
WT	-	99	0	0	0	0
c.68-7T>A	↑ Skipping	92	-24.6	-5.7	2/2	58
c.68-8_68-7delinsAA	↑ Skipping	55	-3.3	-12.4	1/2	97
c.68-3T>A	↑ Skipping	85	-42.6	-4.3	1/2	75
c.68-1G>A*	Δ3p(6nt)	1	-100	-100	2/2	100
c.69T>C	↑ Skipping	92	-11.5	0.0	0/2	7
c.315_316+3delinsGGGTG	↑ Skipping	93	-10.3	-14.0	1/2	100
c.315_316+4delinsGGGTAC	↑ Skipping	37	-20.6	-20.6	2/2	100
c.316G>A	↑ Skipping	34	-16.5	-12.7	2/2	87
c.316G>C	↑ Skipping	58	-4.1	-13.9	1/2	78
c.316+1G>A	↑ Skipping	0	-100	-100	2/2	100
c.316+1G>C	↑ Skipping	0	-100	-59.4	2/2	100
c.316+1G>T	↑ Skipping	0	-100	-49.7	2/2	100
c.316+2T>C	↑ Skipping	4	-100	-0.6	1/2	99
c.316+3A>C	↑ Skipping	3	-34.0	-10.3	2/2	91
c.316+4delA	↑ Skipping	0	-69.1	-29.0	2/2	100
c.316+4A>C	↑ Skipping	34	-15.5	-11.3	2/2	80
c.316+4_316+6delinsCGA	↑ Skipping	5	-26.8	-16.8	2/2	98
c.316+5G>A	↑ Skipping	5	-32.0	-12.7	2/2	96
c.316+5G>C	↑ Skipping	5	-34.0	-13.2	2/2	97
c.316+6T>A	↑ Skipping	26	-6.2	-5.6	1/2	21
c.316+6T>C	↑ Skipping	35	-8.2	-6.0	1/2	25
c.316+6T>G	↑ Skipping	23	-4.1	-5.2	1/2	14
c.68-7delT	No effect	99	8.2	0	2/2	5
c.68-4A>C	No effect	100	-8.2	0	2/2	5
c.316+12A>G	No effect	99	n/a	n/a	n/a	n/a
c.316+13A>G	No effect	100	n/a	n/a	n/a	n/a
c.316+40G>C	No effect	99	n/a	n/a	n/a	n/a
True calls	Positives		15	19	13	21
	Negatives		2	2	2	2
	Total		**17**	**21**	**15**	**23**
False calls	Positives		0	0	0	0
	Negatives		7	3	9	1
	Total		**24**	**24**	**24**	**1**
Sensitivity			68	86	59	95
Specificity			100	100	100	100
Accuracy			71	88	63	96

Table S3. Comparison of experimental results obtained with pCAS2-BRCA2e3 minigenes carrying natural variants mapping to *BRCA2* exon 3 splice sites or to flanking intronic positions with *in silico* data obtained with splice site-dedicated bioinformatics approaches.

Variations *BRCA2* Exon 3 (n = 70)	Effect on splicing	Exon 3 inclusion (%)	QUEPASA (-0.50)	HEXplorer (-14)	SPANR (-0.1)	HAL (-3.4)	LR_{skip} (31.1%)	QUEPASA & HAL	At least 3
WT	-	99	0	0	0	0	0	n/a	n/a
c.74G>T	↑ Skipping	88	-1.65	-24.6	-0.79	1	37.8	1/2	3/4
c.75A>G	↑ Skipping	67	-2.23	-44.8	-1.1	-1.1	51.0	1/2	3/4
c.75A>C	↑ Skipping	63	-2,35	-35,2	-1,78	-1	50,0	1/2	3/4
c.75A>T	↑ Skipping	93	-2,13	-37,3	-1,42	0,5	47.1	1/2	3/4
c.78A>C	↑ Skipping	92	-0,40	-29,5	-2,66	-2	30.3	0/2	2/4
c.79A>G	↑ Skipping	94	-1,04	-5,9	0,16	0	27.2	1/2	1/4
c.90T>C	↑ Skipping	95	1,06	48,6	0,22	1,7	7,2	0/2	0/4
c.92G>A	↑ Skipping	57	-1,69	-29,5	-2,5	-5,1	44,3	2/2	4/4
c.99A>G	↑ Skipping	73	-1,71	-56,0	-0,97	-11,9	54,9	2/2	4/4
c.100G>A*	↑ Skipping Δ3p(6nt)	77	-3,09	-92,4	-1,46	-27,9	83,0	1/2	4/4
c.102A>G	↑ Skipping	51	-1,61	-19,5	-0,94	-1,4	37,4	1/2	3/4
c.107C>T	↑ Skipping	79	-2,36	-140,0	-1,46	-71	94,4	2/2	4/4
c.117T>C	↑ Skipping	92	-0,71	-65,9	-0,28	-12,8	47,4	2/2	4/4
c.120A>C	↑ Skipping	73	-0,61	-52,9	-1	-12,4	42,9	2/2	4/4
c.139G>T	↑ Skipping	91	-1,57	-46,7	-0,53	-1,4	44,5	1/2	3/4
c.140C>T	↑ Skipping	92	-3,34	-112,4	-0,7	-48,7	91,5	2/2	4/4
c.142G>A	↑ Skipping	92	-2,74	-72,6	-2	-25,1	76,1	2/2	4/4
c.143A>G	↑ Skipping	93	-2,64	-57,4	-1,7	-20,1	69,3	2/2	4/4
c.144A>G	↑ Skipping	93	-2,17	-68,5	-1,08	-5	59,6	2/2	4/4
c.145G>T	↑ Skipping	62	-3,69	-144,7	-2,55	-60	96,2	2/2	4/4
c.162C>T	↑ Skipping	84	-1,53	-120,1	-2,32	-23,8	76,8	2/2	4/4
c.165C>T	↑ Skipping	85	-1,88	-103,2	-0,82	-19,2	72,9	2/2	4/4
c.171C>T	↑ Skipping	92	-1,23	-24,8	-0,8	-2,6	35,7	1/2	3/4
c.192T>C	↑ Skipping	89	-1,04	-32,4	-0,05	-15	42,0	2/2	3/4
c.195A>C	↑ Skipping	82	-1,40	-48,8	-0,28	-12,3	49,2	2/2	4/4
c.207A>C	↑ Skipping	88	-0,24	-45,8	-0,35	-11,7	36,3	1/2	3/4
c.220C>T	↑ Skipping	90	-0,31	-62,0	-0,34	-18,9	45,6	1/2	3/4
c.222G>A	↑ Skipping	68	-2,73	-48,3	-1,23	-28,8	71,6	2/2	4/4
c.228A>C	↑ Skipping	76	-1,06	-40,8	-0,47	-6,7	40,4	2/2	4/4
c.231T>G	↑ Skipping	69	1,65	27,2	0,71	1,7	7,2	0/2	0/4
c.231T>C	↑ Skipping	77	-1,51	-21,2	0,23	-6,8	38,8	2/2	1/4
c.252A>G	↑ Skipping	92	-1,28	-58,0	-0,69	-8,7	49,0	2/2	4/4
c.256C>T	↑ Skipping	68	-1,32	-28,8	-0,25	-30,9	53,0	2/2	4/4
c.258G>A	↑ Skipping	92	-2,43	-18,3	-1,17	-28,8	60,9	2/2	4/4
c.258G>T	↑ Skipping	91	-2,00	-24,6	-0,95	-20,3	53,5	2/2	4/4
c.264G>A	↑ Skipping	92	-1,36	27,9	-0,59	-0,9	23,1	1/2	2/4
c.267G>C	↑ Skipping	83	-1,62	-77,7	-1,47	-65,6	84,0	2/2	4/4
c.267G>T	↑ Skipping	94	-0,69	-81,9	-1,42	-14,2	53,6	2/2	4/4
c.270G>A	↑ Skipping	92	-1,75	-34,8	-0,94	-7,4	46,5	2/2	4/4
c.279T>C	↑ Skipping	83	-1,21	-21,2	-0,07	-59,4	65,1	2/2	3/4
c.289G>T	↑ Skipping	-	-2,53	-149,5	-1,23	-38,2	90,5	2/2	4/4
c.291A>G	↑ Skipping	88	-1,42	-8,7	-0,15	-3,5	33,1	2/2	3/4
c.303C>T	↑ Skipping	94	-2,15	-125,5	-1,62	-15,7	78,8	2/2	4/4
c.81A>T	No effect	99	-2,06	-13,8	-0,55	-5,9	42,4	2/2	3/4
c.92G>C	No effect	99	1,60	44,6	0,11	1,9	6,1	0/2	0/4
c.114A>G	No effect	97	-1,37	-81,3	-0,89	-6	55,5	2/2	4/4
c.120A>T	No effect	96	0,26	-50,6	-0,99	-2,6	28,9	0/2	2/4
c.121C>T	No effect	99	0,92	14,8	-0,13	1,2	11,2	0/2	1/4
c.122C>T	No effect	99	0,28	-4,0	-0,28	0,2	17,5	0/2	1/4
c.135A>G	No effect	97	-1,46	-43,3	-0,2	0,2	41,1	1/2	3/4
c.147A>G	No effect	96	-1,03	-29,7	-1,72	-1	35,0	1/2	3/4
c.150T>C	No effect	99	0,39	17,1	0,32	1,4	13,1	0/2	0/4
c.167A>C	No effect	99	0,19	-15,6	-0,14	0,6	20,0	0/2	2/4
c.175C>G	No effect	100	0,37	9,6	-0,09	1,6	14,4	0/2	0/4
c.177A>C	No effect	100	-1,03	-8,2	-0,27	0,2	27,8	1/2	2/4
c.177A>T	No effect	98	-1,32	-50,7	-0,96	-3,5	44,6	2/2	4/4
c.179A>G	No effect	99	-0,67	13,9	-0,07	1,4	19,8	1/2	1/4
c.180C>T	No effect	96	-0,33	-27,9	-0,14	-0,9	27,1	0/2	2/4
c.183A>G	No effect	98	2,15	25,7	1,58	1	5,9	0/2	0/4
c.191C>T	No effect	99	-0,89	-41,3	-0,76	-8,8	40,2	2/2	4/4
c.198A>G	No effect	99	-0,10	9,1	0,85	0,6	16,9	0/2	0/4
c.222G>T	No effect	97	-1,14	-67,2	-0,49	-23,9	58,6	2/2	4/4

c.223G>C	No effect	99	0,31	-3,3	-0,73	0,3	17,4	0/2	1/4
c.237A>C	No effect	99	2,09	55,6	0,31	1,3	4,5	0/2	0/4
c.240A>G	No effect	99	1,26	36,0	0,35	0,9	7,8	0/2	0/4
c.241T>A	No effect	99	0,07	12,1	-0,08	1,2	15,7	0/2	0/4
c.243C>T	No effect	96	-2,54	-131,1	-1,96	-12,6	81,6	2/2	4/4
c.258G>C	No effect	98	-2,14	-20,1	-1,39	-2,8	44,1	1/2	2/4
c.260C>G	No effect	97	-1,04	-10,4	-0,63	-1,5	29,6	1/2	2/4
c.266C>T	No effect	97	0,01	19,2	-0,19	1,5	14,9	0/2	1/4
c.267G>A	No effect	96	-1,46	-29,8	-1,24	-5,1	41,1	2/2	4/4
c.270G>T	No effect	99	-1,75	-34,8	-0,94	-7,4	46,5	2/2	4/4
c.280C>T	No effect	97	0,10	-25,2	0,24	-2,5	23,5	0/2	1/4
c.305A>T	No effect	98	-0,61	-37,1	-1,16	-2,7	33,4	1/2	3/4
True calls		Positives	38	38	37	31	38	28	38
		Negatives	16	17	10	23	19	25	20
		Total	**22**	**55**	**47**	**54**	**57**	**53**	**58**
False calls		Positives	15	14	21	8	12	6	11
		Negatives	5	5	6	12	5	15	5
		Total	**20**	**19**	**27**	**20**	**17**	**21**	**16**
Sensitivity			88	88	86	78	88	65	88
Specificity			52	55	32	74	61	81	65
Accuracy			73	74	64	73	77	72	78

Table S4. Comparison of the experimental results obtained with pCAS2-*BRCA2*e3 minigenes carrying natural *BRCA2* exon 3 variants with *in silico* data obtained with SRE-dedicated bioinformatics approaches.

Partie IV : Discussion

I. Importance des analyses sur l'épissage dans l'interprétation biologique des variations associées à dans les maladies génétiques

L'interprétation biologique et clinique des variations nucléotidiques détectées dans le génome des patients a toujours représenté un défi de la génétique médicale. Cependant, ces dernières années, ce défi a augmenté de façon exponentielle, en particulier depuis l'implémentation, en routine, du séquençage à haut-débit dans les services de diagnostic moléculaire et la découverte concomitante de la grande variabilité du génome humain. L'identification de la mutation à l'origine de la pathologie est en effet essentielle pour le diagnostic moléculaire et le conseil génétique des familles. De plus, la nécessité d'interprétation des VSI a pris davantage d'ampleur avec le développement de nouvelles thérapies ciblées visant à traiter les patients porteurs de mutations pathogènes sur certains gènes (Liu *et al.*, 2015; Vega *et al.*, 2009). Dans ce contexte, la principale mission de notre groupe de recherche est de développer des stratégies visant à faciliter l'interprétation des VSI, en particulier au niveau de l'ARN, en développant des stratégies de criblage de mutations qui affectent l'épissage. Ce processus, à l'origine de la complexité et de la diversité du protéome, est hautement régulé afin d'assurer le contrôle des sites d'épissage utilisés, l'exclusion des pseudo-exons et l'ignorance des sites d'épissage cryptiques selon le type cellulaire et le stade de développement. Dans ce contexte à l'équilibre finement contrôlé, toute variation nucléotidique, qu'elle soit intronique ou exonique, rare ou polymorphique, est susceptible d'affecter l'épissage, soit en modulant la force des sites d'épissage soit en altérant des éléments de régulation. Ainsi, il est possible que de nombreuses VSI que le séquençage parallèle massif a permis d'identifier dans le génome de patients entraînent des défauts d'épissage pouvant conduire à des maladies génétiques ou à des phénotypes particuliers.

Les travaux menés au sein du laboratoire depuis une dizaine d'années, et auxquels j'ai participé pendant ma thèse, ont permis de tester environ 600 variations dans plusieurs gènes et notamment dans les gènes de prédispositions aux cancers les plus fréquents, parmi lesquels les gènes *MMR* (syndrome de Lynch) et *BRCA* (syndrome seins-ovaires), et d'en reclasser près de 25% en mutation d'épissage. Si les sites consensus d'épissage sont les plus fréquemment altérés, de nombreuses variations seraient potentiellement pathogène par modulation d'un élément régulateur d'épissage. En effet, des études mutationnelles exhaustives récemment ciblées sur l'exon 10 de *MLH1* et l'exon 7 de *BRCA2* ont révélé qu'une fraction bien plus importante qu'initialement

estimée affecte des séquences régulatrices de l'épissage (Di Giacomo *et al.*, 2013; Soukarieh *et al.*, 2016). Afin de mieux appréhender la contribution des altérations des éléments régulateurs de l'épissage dans les gènes impliqués dans le syndrome de Lynch et le syndrome seins-ovaires, nous appliquons désormais la stratégie d'analyse systématique des variations répertoriées dans nos exons modèles, et éventuellement dans les régions introniques flanquantes, à l'aide de tests fonctionnels indicateurs d'anomalies d'épissage basés sur l'utilisation de minigènes ou à partir de l'ARN du patient, lorsque disponible. Les travaux accomplis pendant ma thèse ont permis la cartographie mutationnelle des éléments régulateurs de l'épissage dans de nombreux exons des gènes MMR et BRCA, parmi lesquels l'exon 5 de *BRCA1* (article #1), l'exon 5 de *MSH2* (article #1), l'exon 7 de *BRCA2* (article #2), l'exon 7 de *MLH1* (article #3) et l'exon 3 de *BRCA2* (article #4) mais également les exons 9-10, 18 et 21 de *BRCA1* (données non montrées), ainsi que dans d'autres gènes, en particulier l'exon 14 de *MET* et l'exon 10 de *MAPT* (article #1). Les données obtenues pour un total d'environ 1000 variations nous ont permis de confirmer que mutations affectant des éléments régulateurs de l'épissage sont effectivement très fréquentes et par conséquent que toute variation nucléotidique, qu'elle soit intronique ou exonique, est susceptible d'affecter l'épissage de l'ARN et ce, indépendamment de l'impact prédit sur la séquence protéique (ce qui inclue les variations non-sens et *frame-shift*) ou de sa position par rapport aux sites d'épissage. L'ensemble de ces travaux mettent en exergue la nécessité des analyses sur l'épissage de l'ARN dans l'interprétation biologique de variations identifiées dans le cadre du syndrome de Lynch ou du syndrome seins-ovaires, mais également dans toute autre pathologie d'origine génétique.

Dans le but de suivre le rythme actuel de détection de variations par les méthodes de séquençage à haut débit et de répondre à la demande croissante d'analyses de VSI par des « tests minigènes » de la part des laboratoires de diagnostic, il serait important, à l'avenir, de mettre en place une structure de prestation de services (plateforme) dédiée à l'interprétation des variations détectées chez dans le génome de patients évocateurs d'une maladie génétique, en focalisant sur le potentiel impact de ces variations sur l'épissage. Pour ce faire, il serait important d'améliorer les tests fonctionnels basés sur l'utilisation de minigènes pour augmenter leur efficacité et leur débit, tout en réduisant les temps et les coups d'analyses, le but étant de proposer un test « minigène » standardisé reproductible, robuste, fiable, rapide et abordable. Récemment, une plateforme de ce genre a d'ailleurs été développée au sein de l'IGBM (Institut de de Biologie et

Génétique Moléculaire, Valladolid) par l'équipe du Dr Eladio Velasco. Cette plateforme propose : (i) des tests indicateurs d'anomalies d'épissage basés sur l'utilisation du vecteur pSAD, déjà validé pour l'analyse de nombreuses variations génétiques, notamment identifiées dans les gènes *BRCA1*, *BRCA2*, *MLH1*, *COL1A1*, *SERPINA1*, *CHD7* (Acedo *et al.*, 2012a, 2012b, 2015; Fraile-Bethencourt *et al.*, 2017; Lara *et al.*, 2014; Sanz *et al.*, 2010; Villate *et al.*, 2018), (ii) des analyses sur ARN de patients, à partir de lignées lymphoblastoïdes ou d'échantillons de sang, et (iii) des prédictions bioinformatiques axées sur l'épissage.

II. Le minigène, un outil essentiel des tests fonctionnels indicateurs d'anomalies d'épissage

Pour des raisons évidentes, la méthode la plus appropriée pour identifier les anomalies d'épissage est basée sur l'analyse comparative des profils d'épissage des transcrits exprimés dans les tissus du patient porteur de la variation avec ceux obtenus chez des individus témoins. Cependant, le tissu relevant étant rarement disponible, et alternativement, les laboratoires de diagnostic ne disposant pas toujours d'échantillons d'ARN ou de lignées cellulaires lymphoblastoïdes dérivés du sang des patients, le « test minigène » reposant sur l'utilisation de l'ADN génomique du patient constitue une approche alternative pour évaluer les conséquences fonctionnelles des variations sur l'épissage. En effet, le recours au minigène nous a permis d'évaluer l'effet sur l'épissage d'environ 1000 variations et ce, sans aucun besoin du matériel biologique de patients pour la plupart, puisque plus de 95% des minigènes porteurs des variations étudiées au cours de ces travaux de thèse ont été préparées par mutagenèse dirigée. Malgré l'aspect artificiel de ce système, nous avons pu valider cette approche par de nombreux travaux, y compris ceux effectués sur les exons 3 et 7 de *BRCA2* (articles #2 et 4) démontrant la très bonne concordance entre les données obtenues à partir des tests fonctionnels basés sur l'utilisation de minigènes et celles obtenues par l'analyse de l'ARN des patients (Bonnet *et al.*, 2008; Gaildrat *et al.*, 2012; Houdayer *et al.*, 2012; Tournier *et al.*, 2008).

Malgré le développement de tests fonctionnels basé sur la modification du génome à saturation et combinant des analyses sur l'ARN et la protéine très prometteurs (Chapitre VI - Section 4), le « test minigène » reste une approche incontournable. En effet, bien qu'une quantité incroyable de VUS aient d'ors et déjà été analysées en parallèle et très rapidement (4000 variations

dans *BRCA1*), cet essai fonctionnel se limite pour l'instant à l'analyse d'un nombre très restreint de gènes, essentiels à la viabilité de la lignée cellulaire HAP1, parmi lesquels, les gènes de la voie de réparation par recombinaison homologue, en particulier *BRCA1/2*, *RAD51C/D* et *PALB2* (Findlay *et al.*, 2018). Il ne peut donc pas, pour l'instant, être appliqué tel quel à n'importe quel gène, contrairement au test minigène et aux prédictions bioinformatiques axées sur l'épissage, qui peuvent être appliqués de manière universelle à l'analyse de n'importe qu'elle variation situées en dehors des exons terminaux ou de grande taille. De plus, l'implémentation du test minigène dans des MAVEs vont permettre son automatisation et de facto de réduire les coûts et les temps d'analyse, afin de proposer un test fonctionnel aussi rapide et efficaces que ceux basé sur la modification du génome, mais largement moins couteux.

L'une des limitations majeures des analyses des profils d'épissage à partir d'ARN dérivés du sang des patients est la dégradation possible, par le NMD, des transcrits contenant des codons stop prématurés, rendant ces transcrits aberrants difficilement détectables et pouvant conduire à une mauvaise interprétation quant à l'effet de la variation étudiée sur l'épissage. Ce problème peut être résolu, dans certains cas, en en établissant des lignées cellulaires stables à partir du sang du patient et en les traitant avec des inhibiteurs du NMD. Cependant, cela représente un investissement considérable en temps et en ressources, qu'il serait difficile d'implémenter dans des analyses de routine à haut-débit (Baralle *et al.*, 2009; Spurdle *et al.*, 2008). L'utilisation de minigène représente donc une alternative intéressante, en particulier lorsque les transcrits générés à partir de certains minigènes, comme le minigène pCAS2 développé par notre unité, sont résistants à la dégradation par le système NMD. Cependant, nous avons utilisé dans le cadre des travaux menés sur *MAPT* exon 10 (article #1), le minigène pSPL3m, qui contrairement au pCAS2, génère des transcrits sensibles à la dégradation par le NMD, nécessitant l'utilisation d'inhibiteurs du NMD. Pour pallier à cette limitation, nous avons entrepris de modifier le vecteur pSPL3m par mutagenèse dirigée en inactivant le codon d'initiation de la traduction afin de produire des transcrits résistants au NMD. Après validation, la version modifiée du pSPL3m (pSPL3mK) a pu être utilisée pour évaluer l'effet sur l'épissage des variations de l'exon 10 de *MAPT*, sans utilisation d'inhibiteurs du NMD (Article #1). De plus, ce vecteur comporte (i) un site donneur d'épissage cryptique dans l'intron présent entre les exons A et B du minigène, pouvant être activé lorsqu'un site activateur d'épissage cryptique est également présent dans l'intron flanquant l'exon d'intérêt (Soukarieh *et al.*, 2016) et (ii) un exon cryptique de 117 nucléotides dans l'intron du pSPL3m, lequel est dérivé du HIV,

pouvant être inclus sous certaines conditions (Burn *et al.*, 1995), qui limitent l'efficacité du test *ex vivo* d'épissage basé sur l'utilisation du minigène pSPL3m. Ainsi, nous avons également inactivé le site donneur d'épissage cryptique puis délété l'exon cryptique de façon à améliorer l'efficacité du vecteur minigène. Nous souhaitons dans l'avenir valoriser ce travail en (i) analysant les profils d'épissage des minigènes sauvages (i.e., l'exon 12 de *BRCA2*, l'exon 10 de *MLH1* et l'exon 15 de *MSH2)* associés à l'utilisation du site donneur d'épissage cryptique et éventuellement à l'inclusion de l'exon cryptique du pPSL3m et (ii) ré-analysant des variations qui auraient été analysées auparavant avec le pSPL3m et qui pourraient éventuellement être associées à une interprétation erronée du défaut d'épissage induit par la variation de par la dégradation, par le NMD, de transcrits aberrant porteurs d'un PTC.

Nous disposons au sein du laboratoire de vecteurs minigènes différents, qui nous permettent d'adapter le test fonctionnel indicateur d'anomalies d'épissage selon l'exon d'intérêt : (i) le minigène pCAS2 est utilisé en routine dans le laboratoire, (ii) le minigène pSPL3m, utilisé eb deuxième intention pour l'analyse d'exons particuliers, dont le profil d'épissage physiologique n'est pas correctement reproduit dans le minigène pCAS2 et (iii) le minigène pB1, reçu généreusement de la part de Diana Baralle (Université de Southampton, Grande-Bretagne) et spécifiquement utilisé pour l'analyse des exons 9-11 de *BRCA1*. En principe, nous choisissons le minigène dans lequel l'exon à tester se comporte de la manière la plus proche de l'état physiologique, c'est-à-dire le minigène qui reproduit le mieux le profil d'épissage physiologique naturel de l'exon d'intérêt observé à partir de l'analyse d'échantillons d'ARN dérivé du sang d'individus contrôles. En effet, le choix du minigène pSPL3mK pour évaluer l'effet sur l'épissage des variations de l'exon 10 de *MAPT* repose sur le fait que, à l'état sauvage, l'exon 10 est partiellement exclu dans le contexte du minigène pSPL3m, reflétant le profil d'épissage alternatif de cet exon dans le cerveau adulte à l'état physiologique, tandis qu'il était totalement inclus dans le minigène pCAS2. De même, le minigène pB1, est spécifiquement utilisé pour analyser les variations identifiées dans les exons 9-11 de *BRCA1* car il présente la particularité de contenir les exons 8-12 du gène *BRCA1* permettant de reproduire l'épissage alternatif des exons 9-11 de *BRCA1*. Dans ce contexte, il est alors possible d'évaluer l'impact des variations identifiées au niveau des exons 9-10 de *BRCA1* et des régions introniques flanquantes sur l'épissage combiné de ces exons, ce qui n'était pas possible avec le minigène pCAS2, dans lequel on analysait l'épissage

de l'exon 9 ou de l'exon 10 seul, selon que la variation était localisée dans l'exon 9 ou 10, respectivement.

Le choix du minigène à utiliser est d'une importance capitale afin de proposer une interprétation la plus correcte possible de l'effet de la variation analysée sur l'épissage. En effet, il a été récemment démontré, à l'aide de données génétiques et fonctionnelles d'épissage, la relevance clinique des transcrits alternatifs physiologiques en phase Δ9-10 qui pourraient être associés à un mécanisme de compensation de certaines variations supposées pathogènes. Ces variations pourraient augmenter le saut en phase des exons 9-10, apparemment codant pour une séquence non-essentielle à la fonction de la protéine BRCA1 (de la Hoya *et al.*, 2016). Il est donc essentiel d'évaluer l'effet de ces variations sur l'épissage combiné des exons 9 et 10. Ces travaux démontrent ainsi la nécessité de prendre en considération les effets éventuels des variations sur l'épissage alternatif des exons étudiés lors du design des analyses des transcrits. Dans ce contexte, il est possible que certaines des variations analysées dans et autour de l'exon 7 de BRCA2 pour leur effet sur l'épissage de cet exon, puisse également être associée à un saut alternatif en phase des exons 4-7. Cependant, à l'heure actuelle, il n'existe pas, à notre connaissance, de minigène capable d'analyser l'effet sur l'épissage combiné des exons 4-7 de *BRCA2*. Par conséquent, de façon similaire au minigène pB1 contenant les exons 8-12 de *BRCA1*, nous avons construit le minigène pCAS2-BRCA2e2-9 (appelé pB2) contenant les exons 2-9 de BRCA2 ainsi qu'environ 200-250 nucléotides des régions introniques flanquantes. Ce minigène pourrait permettre l'étude de l'épissage alternatif de *BRCA2* au niveau des exons 3-7, ce qui n'était jusqu'ici pas possible avec le vecteur minigène pCAS2-BRCA2e7. Nous envisageons maintenant de valider l'utilisation de ce minigène par comparaison de son profil d'épissage avec celui des transcrits *BRCA2* exprimés dans le sang d'individus témoins et de patients lorsque disponible. Nous pourrons ensuite poursuivre notre étude portant sur l'exon 7 de *BRCA2* et de l'étendre aux exons 4-6 en amont, afin d'identifier des variations potentiellement hypomorphes induisant un saut des exons 4-7. De telles variations pourraient éventuellement être associées à un risque intermédiaire de développer un cancer du sein et/ou de l'ovaire.

III. Stratégie de stratification des variations pour des études fonctionnelles d'épissage par des outils de prédictions bioinformatiques

Les données fonctionnelles d'épissage générées au laboratoire, y compris pendant ces travaux de thèse ont révélé qu'une proportion importante des variations nucléotidiques identifiées dans le génome de patients évocateurs d'une pathologie d'origine génétique pourrait altérer l'épissage de l'ARN. Toutefois, étant donné le nombre excessivement important de variations actuellement détectés par séquençage à haut débit, il n'est pas envisageable à ce jour que chacune de ces variations ne fassent l'objet d'une analyse fonctionnelle permettant d'évaluer leur impact sur l'épissage. Il est donc indispensable de sélectionner de manière rationnelle les variations à analyser en priorité à l'aide de tests fonctionnels indicateurs d'anomalies d'épissage. Aujourd'hui, ces stratégies de stratification des VSI s'appuient principalement sur l'utilisation d'outils bioinformatiques de prédiction des altérations de sites consensus d'épissage. En effet, de nombreuses études, y compris celles menées sur les exons 7 des gènes *BRCA2* (article #2) et *MLH1* (article #3) ont démontré la grande fiabilité de ce type d'outils qui sont désormais couramment utilisés dans les laboratoires de diagnostic pour identifier des mutations d'épissage, surtout depuis leur implémentation dans l'interface Alamut Visual (Houdayer *et al.*, 2012; Leman *et al.*, 2018; Moles-Fernández *et al.*, 2018; Spurdle *et al.*, 2008; Théry *et al.*, 2011; Tournier *et al.*, 2008). En revanche, les éléments de régulation, encore assez peu caractérisés, manquaient, jusqu'ici, d'outils de prédiction suffisamment performants.

Récemment, une étude pilote menée au laboratoire a révélé, à partir d'un jeu de données de 154 variations identifiées dans 5 exons modèles (*MLH1* exon 10, *BRCA2* exon 7, *CFTR* exon 10, *NF1* exon 37 et *BRCA1* exon 6), les bonnes performances de deux des trois nouvelles approches (QUEPASA et HEXplorer mais pas SPANR) axées sur les altérations des séquences exoniques régulatrices de l'épissage (ESR) évaluées. Nous avons étendu ici l'évaluation du caractère prédictif de ces trois nouvelles approches (QUEPASA, HEXplorer et SPANR) et d'une approche additionnelle (HAL) par une étude comparative à large échelle des scores générés par ces approches avec des données expérimentales obtenues pour un total d'environ 1200 variations exoniques. Nos travaux ont ainsi démontré la fiabilité de l'ensemble de ces approches, utilisées seules ou en combinaison, y compris de SPANR après optimisation du seuil de décision, pour la prédiction des altérations des ESR conduisant à une augmentation du saut ou de l'inclusion. De plus, des données

préliminaires suggèrent que ce type de méthodes pourront également être utiles pour prédire la création d'un pseudoexon. Ces travaux nous ont permis de proposer des recommandations quant à l'utilisation de ces approches en tant qu'outils de filtration pour prioriser les variations potentiellement splicéogéniques à analyser dans des tests fonctionnels axés sur l'épissage, en proposant notamment des seuils de décision. Bien que nous ayons initié les analyses sur la combinaison des outils de prédiction axés sur les altérations d'ESR, avec des résultats très promoteurs (augmentation du pouvoir prédictif comparativement aux approches utilisées seules), nous considérons qu'il serait maintenant important de développer des approches d'apprentissage automatique (*machine learning*) dans le but de proposer et d'évaluer des métaprédicteurs. Ces métaprédicteurs, émanant de la combinaison de plusieurs outils de prédiction des altérations d'ESR pourront être évalués à l'aide notamment de modèles de *random forest* et de *naïve voting method*, de façon similaire aux travaux menés sur les outils de prédiction des altérations faux-sens dans les gènes (Hart *et al.*, 2018).

Néanmoins, une analyse mutationnelle exhaustive ciblée sur l'exon 7 de *MLH1* (article #3), a mis en évidence l'échec apparent de ces approches, pourtant validées par des études menées sur l'exon 7 de *BRCA2* (article #2), l'exon 10 de *MAPT* et l'exon 5 de *MSH2* (article #1), laissant suggérer que ces méthodes pourraient ne pas s'appliquer de manière équivalente à tous les exons et/ou à tous les gènes. En effet, nous avons montré que cet exon était doté de caractéristiques particulières, i.e. de sites d'épissage remarquablement forts, lui conférant une résistance totale aux mutations de régulation d'épissage. Par une stratégie d'affaiblissement des sites d'épissage, mous avons également montré que l'échec apparent des outils de prédictions d'altérations d'ESR était dû, au moins en partie, à la force particulièrement élevée des sites d'épissage. L'ensemble de ces données contribuent à mieux déterminer les limitations et les applications de ces outils bioinformatiques, ce qui pourra contribuer à leur amélioration. De manière intéressante, des données additionnelles obtenues sur l'exon 21 de *BRCA1* (données non montrées), présentant des caractéristiques similaires à celles de l'exon 7 de *MLH1*, i.e. une force intrinsèque des sites d'épissage particulièrement élevée et une petite taille, ont révélé que cet exon semblait également être très résistant aux mutations de régulation d'épissage. En effet, seules 7 des 73 variations analysées (10%) à l'aide du minigène pCAS2 induisent des défauts d'épissage très partiels de l'exon 21 (seulement 5-10% de saut d'exon).

Il serait maintenant intéressant de réaliser des études mutationnelles sur d'autres exons modèles, identifiés à partir de projets de recherche et/ou du diagnostic, représentant des cas de discordances avec les outils de prédiction d'altérations d'ESR. Ces travaux, comme ceux menés sur l'exon 7 de *MLH1*, pourraient permettre d'identifier d'autres caractéristiques particulières ou des variables qui influencent l'utilisation des sites d'épissage et qui pourraient ainsi expliquer les discordances observées entre les données *in silico* et fonctionnelles d'épissage, parmi lesquelles (i) la richesse de l'exon en ESR (Haque *et al.*, 2010; Tammaro *et al.*, 2014), (ii) la densité de sites cryptiques présents le long de l'exon et des séquences introniques flanquantes (Brillen *et al.*, 2017), (iii) l'influence de la taille des exons et des séquences introniques flanquantes (Berget, 1995; Fox-Walsh *et al.*, 2005), (iv) la présence d'éléments régulateurs de l'épissage dans les séquences introniques à proximité de l'exon (Gao *et al.*, 2007; Kashima *et al.*, 2007), (v) les caractéristiques du point de branchement, (vi) la présence de structure secondaire (Buratti and Baralle, 2004; D'Souza and Schellenberg, 2000; Singh *et al.*, 2007), (vii) le degré de conservation (Savisaar and Hurst, 2018; Wainberg *et al.*, 2016), (viii) le caractère alternatif de certains exons, ou encore (ix) la densité en nucléosome et le taux de transcription (Aissat *et al.*, 2013). Bien souvent, ces variables qui affectent l'effet des variations sur l'épissage ou influence la susceptibilité d'un exon aux mutations ESR ne sont pas prises en compte par la plupart des approches de prédictions, car elles sont difficiles à prédire (pour revue : Grodecká *et al.*, 2017). Ces variables pourraient alors expliquer les bonnes performances obtenues sur certains exons et les mauvaises obtenues sur d'autres exons. L'ensemble de ces données soulignent la nécessité de développer des approches de prédiction plus complexes, intégrant ces variables dans le calcul des scores de prédiction d'altérations d'ESR afin de contribuer à l'amélioration les algorithmes préexistants. Compte tenu du jeu de données conséquent constitué au laboratoire par les travaux de recherches, les analyses diagnostiques ainsi que les analyses bibliographiques (n=3000 variations caractérisées pour leur effet sur l'épissage), nous souhaitons maintenant utiliser des approches d'apprentissage automatique afin de développer un modèle de prédiction des altérations de l'épissage, basées sur les approches de prédictions des altérations des ESR prometteuses mais intégrant pleinement l'ensemble des caractéristiques qui influencent l'utilisation des signaux d'épissage. Des analyses similaires, auxquelles nous participons, sont actuellement menées par Sophie Krieger et Raphaël Leman (Inserm U1245, Centre François Baclesse) sur les approches de prédiction des altérations des signaux d'épissage. Si initialement, ces travaux étaient focalisés sur les altérations des sites

consensus d'épissage (Leman *et al.*, 2018), ils sont maintenant étendus aux défauts d'épissage associés à la création d'un site d'épissage de novo et/ou l'utilisation d'un site d'épissage cryptique ainsi qu'aux altérations des points de branchement.

L'ensemble de ces études soulignent l'importance de combiner non seulement les approches bioinformatiques de prédiction des altérations de l'épissage entre elles, mais également avec les tests fonctionnels d'épissage, afin d'établir des stratégies de criblage de mutations splicéogéniques efficaces. Nous espérons que ces travaux contribuent à mieux déterminer les mécanismes et signaux qui régissent l'épissage, et qu'ils puisse aider, avec le développement de l'intelligence artificielle au service de biologie, au développement d'un outil unique, intégrant les codes de l'épissage et capables de prédire toute altération de ce code, ce qui pourrait avoir d'importantes retombées pour l'interprétation et la classification des nombreux SNV identifiés lors du dépistage de maladies génétiques.

IV. Vers le développement des approches fonctionnelles combinant les analyses ARN-protéines

En dépit des avancées réalisées dans le criblage des mutations d'épissage, l'évaluation de la pathogénicité des mutations associées à des défauts d'épissage reste complexe, en particulier lorsque celles-ci conduisent à des anomalies d'épissage en phase et/ou partielles. En effet, si la plupart des mutations provoquent des défauts d'épissage hors phase, permettant de classer ces variations comme pathogènes si absence de transcrits alternatifs pouvant compenser la perte du transcrit pleine longueur, de nombreuses VSI entraînent des modifications en phase et/ou partielles. Notamment, un nombre important de mutations est responsable d'un saut en phase, partiel ou total, de l'exon 3 de *BRCA2* (Δ3). S'il a été démontré, à l'aide de données fonctionnelles et génétiques, que les variations à l'origine du saut total de l'exon 3 sont délétères car elles conduisent à la délétion du domaine d'interaction avec la protéine PALB2, les conséquences des anomalies d'épissage Δ3 partielles sur la fonctionnalité de la protéine BRCA2 résultante demeurent inconnues et de telles variations restent donc classées comme VSI et ne peuvent être utilisées au titre du diagnostic.

Afin d'évaluer l'effet combiné de ces variations sur l'épissage de l'ARN et sur la fonctionnalité de la protéine dans un contexte génomique et cellulaire proche du contexte naturel,

nous avons mis à profit un nouvel essai fonctionnel lors d'une mobilité effectuée dans le laboratoire du Dr Shyam Sharan (NCI/CCR, NIH, Frederick, USA). Il s'agit d'un test de complémentation basé sur l'utilisation de cellules souches embryonnaires de souris (mESC) visant à évaluer la capacité de transgènes BRCA1/2 humains mutés (BAC contenant le gène *BRCA1* ou *BRCA2* humain entier) à restaurer la fonction de *BRCA1* ou *BRCA2,* inactivé de façon conditionnelle (Chapitre VI – Section 3). Contrairement aux essais fonctionnels protéiques reposant sur l'utilisation de vecteur d'expression portant uniquement la région codante du gène (ADNc, ADN complémentaire), ce système permet d'analyser les conséquences fonctionnelles des variations qui pourraient affecter l'épissage, et plus largement, de n'importe quelle variation, incluant celles localisées sur les séquences non-codantes, i.e. introniques, promotrices et non traduites (3' et 5' UTR). En outre, ce système étant monoallélique, comme le test indicateur d'anomalies d'épissage basé sur l'utilisation de minigènes, il permet d'établir une relation de causalité directe entre le défaut d'épissage observé, la fonctionnalité de la protéine et la variation analysée. En outre, nous avons montré que ce système reproduit les défauts d'épissage observés à l'aide des tests fonctionnels d'épissage, basés sur l'utilisation du minigène pCAS2-BRCA2e3 ou sur l'analyse du matériel biologique de patient (article #4 et données non montrées). De plus, l'expression du transgène étant soumise au promoteur naturel, ce système permet de garantir un niveau d'expression proche du contexte naturel, contrairement aux approches utilisant le cDNA souvent associées à une surexpression du transgène porteur de la variation à l'aide d'un promoteur hétérologue fort. Cependant, les protéines BRCA exerçant de nombreuses fonctions dans l'organisme et particulièrement la protéine BRCA1 pléiotrope, il n'existe donc pas de test fonctionnel universel permettant d'évaluer l'ensemble des fonctions des protéines BRCA. Toutefois, la calibration des essais fonctionnels basés sur la recombinaison homologue et la sensibilité aux agents génotoxiques à partir de variations dont la pathogénicité ou la neutralité ont été déterminées grâce à des données génétiques, familiales et cliniques, ont permis la validation de cet essai et son utilisation comme un outil fonctionnel d'aide à l'interprétation des variations génétiques identifiées dans le gène *BRCA2* (Mesman *et al.,* 2018).

Il serait intéressant de mettre à profit cet essai fonctionnel pour l'étude des variations à l'origine d'anomalies d'épissage en phase, en particulier celles localisées dans les exons 4-7 et 12 de *BRCA2* ainsi que dans les exons 9-10 de *BRCA1* et pour lesquelles il a été suggéré l'existence possible d'un mécanisme de correction ou d'un effet protecteur (Introduction – Chapitre 8), et

également à l'études des petites insertions/délétions en phase. D'ailleurs, des travaux sur l'exon 12 de BRCA2 sont actuellement menés dans notre laboratoire, en collaboration avec le Dr Maaike Vreeswijk (Département de génétique humaine, Université de Leiden, Pays-Bas), en ce qui concerne les tests avec des cellules mES. De plus, nous collectons actuellement, en collaboration avec le Groupe Génétique et Cancer, des informations génétiques, cliniques et familiale, sur les variations localisées au niveau de ces exons et des séquences introniques flanquantes et pour lesquelles nous disposons de données fonctionnelles d'épissage, soit à partir des « tests minigènes » soit à partir de l'étude de l'ARN de patients. Ces travaux devraient contribuer à l'interprétation biologique et clinique des variations et à la réévaluation potentielle du caractère pathogène de certaines mutations, avec des conséquences importantes pour le diagnostic moléculaire et la prise en charge médicale des patients et de leurs apparentés.

V. Importance de la caractérisation des anomalies d'épissage dans l'interprétation biologique des variations et la prise en charge des patients et de leurs apparentés

Dans le contexte actuel de médecine personnalisée, l'identification de la mutation à l'origine de la pathologie est essentielle. Si la contribution première concerne le diagnostic moléculaire et le conseil génétique des familles, dans certaines maladies génétiques, l'identification de la mutation a potentiellement un impact sur les stratégies thérapeutiques envisagées. En effet, l'identification de la mutation pathogène conditionne aujourd'hui l'accès à de nouvelles thérapies, notamment des thérapies ciblées très prometteuses telles que celles basées sur les inhibiteurs de PARPs dans le contexte du syndrome seins-ovaires ou des stratégies d'immunothérapies personnalisées dans le syndrome de Lynch. De plus, l'identification de la mutation pathogène contribue à l'inverse à éviter à l'inverse l'administration de traitements anti-cancéreux inefficaces ou l'apparition de phénomène de résistance à certains traitements anti-cancéreux classiquement utilisés, tels que les agents méthylants, les sels de platine et les fluoropyrimidines dans le syndrome de Lynch.

Ces dernières décennies des stratégies thérapeutiques basées sur la modulation ou la correction des anomalies de l'épissage. La faisabilité d'une thérapie basée sur la correction de l'épissage dépend bien évidemment des caractéristiques de la maladie, de son mode de transmission, du stade de développement qui est affecté, de l'âge de début, du tissu atteint et de

son accessibilité au traitement. Ces stratégies semblent prometteuses pour la correction d'anomalies d'épissage spécifiques, à l'origine primaire des maladies monogéniques et particulièrement des désordres neuromusculaires (Lee and Abdel-Wahab, 2016). En effet, ces stratégies ont été efficacement appliquées à la correction de défauts d'épissage causés par certaines mutations dans les gènes codant la dystrophine (*DMD,* approches de « saut d'exon ») ou la protéine de survie des motoneurones (*SMN2,* approches de « réinclusion de l'exon ») avec le développement d'essais cliniques et de deux médicaments, l'Eteplirsen et le Nusinersen, ayant récemment obtenu une autorisation de mise sur le marché dans le traitement de la myopathie de Duchenne et l'amyotrophie spinale, respectivement (Lee and Abdel-Wahab, 2016). Dans le cas des prédispositions génétiques aux cancers, il n'existe pas, à ma connaissance, de thérapies basées sur la correction de l'épissage en clinique. En effet, la correction de défauts constitutionnels d'épissage pourrait avoir un potentiel thérapeutique très limité car ces altérations ne sont que le début d'une instabilité génétique qui induit l'accumulation de nombreuses mutations somatiques, dont une partie est impliquée directement dans le développement tumoral. De plus, ces stratégies pourraient être trop risquées, les patients porteurs d'une mutation pathogène à l'état hétérozygote n'étant malades.

L'identification et la caractérisation du défaut constitutionnel chez des patients évocateurs d'une prédisposition au cancer est également essentielle pour appréhender les relations génotypes-phénotypes. En effet, il a été montré dans le gène de prédisposition à la dystrophie musculaire de Duchenne, que des mutations non-sens ou *frame* responsables de l'exclusion d'un ou plusieurs de exons permettent de supprimer des codons stop prématurés dans le cas de mutations non-sens ou de restaurer la phase dans le cadre de mutations *frame-shift*, conduisant ainsi la production d'une protéine comportant d'une délétion interne, mais fonctionnelle au moins en partie. Ce type de mutation est alors associée à l'existence d'un phénotype atténué particulier correspondant à une forme moins sévère de la dystrophie musculaire de Duchenne, appelée dystrophie musculaire de Becker. Cela nécessite, au préalable, de caractériser de manière extensive le profil d'épissage du gène d'intérêt et l'impact fonctionnel des isoformes sur l'activité de la protéine afin d'identifier des exons potentiellement non essentiels à l'activité de la protéine. Dans le contexte du syndrome seins-ovaires, l'analyse extensive des profils d'épissage des gènes *BRCA1* et *BRCA2*, ont rapporté l'existence de nombreux transcrits alternatifs comportant des délétions internes en phase dans la région codante et pouvant générer des isoformes fonctionnelles. De plus, des analyses

mutationnelles ciblées sur certains de ces exons alternatifs ont permis d'identifier que certaines mutations d'épissage entrainent des modifications en phase de la protéine en dehors d'une région codant pour un domaine fonctionnel protéique et/ou ne diminuent pas l'expression de transcrits alternatifs. Les isoformes résultantes, notamment BRCA1 Δ9-10, BRCA2 Δ4-7 et BRCA2 Δ12, pourraient être fonctionnelles, au moins en partie. Des travaux sont actuellement en cours afin de confirmer l'existence possible d'un mécanisme de correction ou de protection par les isoformes BRCA2 Δ4-7 et Δ12 et BRCA1 Δ9-10 à l'aide, d'une part, d'analyses fonctionnelles d'épissage et/ou protéiques et, d'autre part, à l'aide d'analyses multifactorielles, basées notamment sur la collecte des données génétiques, cliniques, tumorales et familiales des patients porteurs. L'ensemble de ces données devraient permettre de statuer sur le caractère hypomorphe de certaines de ces variations et d'appréhender l'existence possible d'une variabilité de l'expression clinique du syndrome seins-ovaires observée chez les patients porteurs, notamment une modification du risque de développer un cancer, de la sévérité de la maladie et de la pénétrance. Ces travaux devraient ainsi contribuer à mieux comprendre relations génotypes-phénotypes et dans l'avenir, d'adapter la prise en charge des patients et de leurs apparentés.

Afin de mieux appréhender les relations génotypes-phénotypes, il est aujourd'hui capital de caractériser les anomalies d'épissage de manière quantitative ou semi quantitative. A l'heure actuelle, la majorité des travaux analysant l'impact des mutations sur l'épissage reposent uniquement sur des analyses qualitatives, consistant à déterminer si la variation altère ou non l'épissage, et le cas échéant, la nature (saut d'exon, délétion d'une partie de l'exon ou rétention d'une partie de l'intron) et éventuellement l'importance (total ou partiel) du défaut d'épissage observé mais ne permettent pas d'appréhender la sévérité. Les travaux menés pendant ma thèse ont permis de quantifier l'effet sur l'épissage de plus de 1000 variations à l'aide d'un test fonctionnel d'épissage (« test minigène » ou ARN de patient) couplé à une analyse semi-quantitative par RT-PCR fluorescente et électrophorèse capillaire. Or, déterminer de la sévérité des défauts d'épissage observé sur un exon donné est d'une importance cruciale pour la stratification bio-informatiques et l'interprétation biologique des variations. D'une part, la quantification des défauts d'épissage va contribuer à la calibration des outils de prédictions bio-informatiques, en corrélant la valeur du score généré par l'outil de prédiction à l'importance du saut d'exon, permettant ainsi de prédire la sévérité du défaut d'épissage et d'orienter les tests fonctionnels vers des variations potentielle à effet fort. D'autre part, la quantification des défauts d'épissage va contribuer à la calibration de la

pathogénicité en corrélant la sévérité du défaut d'épissage observé avec la fonction de la protéine et le phénotype observé notamment la sévérité de la maladie. Nous avons montré grâce aux travaux portant sur l'exon 3 de *BRCA2*, qu'une réduction substantielle des taux de transcrits *BRCA2* pleine longueur (environ 70%) ne réduisait pas ni la viabilité ni la fonctionnalité de la protéine BRCA2 dans le modèle de cellules mES (article #4). De même, il a été montré qu'une réduction substantielle des transcrits *BRCA1* pleine longueur concomitante avec un taux résiduel suffisant de transcrits alternatifs Δ9-10 (~20%) ne serait pas associée à une risque seraient risque élevé de développer un cancer (de la Hoya *et al.*, 2016). Il est possible que des mutations d'épissage à effet partiel plus ou moins important conduisant à une réduction plus ou moins substantielle des taux de transcrits et de la fonctionnalité des protéines BRCA, ce qui pourrait conduire à une variabilité du phénotype observé. De telles variations pourraient être associées à un risque intermédiaire de développer un cancer. A l'avenir, il serait donc important de calibrer, probablement par exon ou par domaine fonctionnel, la pathogénicité afin de déterminer des seuils de pathogénicité discriminant entre des variations à effets partiels faibles associés à une protéine BRCA fonctionnelle et un risque de cancer faible ou modéré et des effets partiels forts associés à une protéine non fonctionnelle et un risque de cancer élevé.

VI. Contribution des collaborations et des consortia dans l'interprétation et la classification des variations génétiques

Devant les nombreuses VSI détectées dans le cadre du diagnostic moléculaire, il paraît aujourd'hui essentiel de rassembler, au niveau national et international, les efforts entre cliniciens et scientifiques pour parvenir à la classification clinique et l'interprétation de l'effet biologique de ces variations. En effet, bien que la combinaison des données *in silico* et fonctionnelles d'épissage apporte sa contribution à l'interprétation de l'impact de ces variations sur l'épissage, ces données aboutissent rarement, à elles seules, à la classification de ces variations selon leur pathogénicité. Le domaine de l'interprétation des variations génétiques fait appel, en réalité, à une expertise pluridisciplinaire rassemblant des compétentes en bioinformatique (données de prédictions d'épissage et protéiques), en biostatistiques (études d'association), en biologie moléculaire et cellulaire (données fonctionnelles), en génétique humaine (données de population, de co-

ségrégation, de co-occurrence, données familiales) et en clinique (données tumorales et phénotypiques).

Plusieurs consortia et groupes de travail, avec lesquels nous collaborons pour certains, ont été créés aux niveaux national et international dans le but de partager les expertises et compétences dans les nombreuses spécialités touchant à l'interprétation des variations et de favoriser les interactions entre cliniciens et scientifiques. Ces consortia ont également permis de standardiser et harmoniser les règles de classification des variations. En effet, bien que des bases de données aient été créées dans le but de rassembler le maximum d'informations collectées sur les variations, ces bases de données sont très nombreuses et l'interprétation biologique attribuée à une variation donnée est parfois conflictuelle voire contradictoire. De ce fait, certains de ces consortia ont élaboré des directives encadrant l'interprétation des variations de séquence identifiées dans le génome de patients, directives généralisées à toute maladie génétique comme celles de l'ACMG-AMP (Richards *et al.*, 2015) ou limitées à certaines maladies génétiques particulières, comme celles d'InSiGHT (syndrome de Lynch, Thompson *et al.*, 2014) ou d'ENIGMA (syndrome seins-ovaires, Spurdle *et al.*, 2012).

Pourtant, l'interprétation des variations génétiques restent aujourd'hui très délicate, en particulier parce que l'accès aux informations liées aux patients reste souvent limité, mettant en exergue la nécessité absolue de partager les données collectées sur les variations, patients et familles par les laboratoires de diagnostic et de recherche et d'intégrer de manière globale l'ensemble de ces données. Notre collaboration avec le consortium ENIGMA et le Groupe Génétique et Cancer au sein des études accomplies notamment sur les exons 3 et 7 de *BRCA2* illustre parfaitement la nécessité de partages de données. En effet, ces collaborations nous ont permis d'augmenter les chances d'obtenir des informations génétiques, familiales et cliniques sur les patients porteurs de variations rares voire uniques. Pour un petit nombre de variations, nous espérons ainsi disposer de suffisamment d'informations pour conclure sur la pathogénicité. Dans cette optique, le *Broad Institute of MIT and Harvard*, *l'Ontario Institute for Cancer Research* et le *Wellcome Trust Sanger Institute*, auxquels se sont jointes plus de 500 organisations internationales publiques et privées (dont l'Université de *California Santa Cruz*, l'Université de Cambridge, le *National Institutes of Health*, l'Institut National du Cancer et France Génomique), finance, depuis 2013, l'organisation non gouvernementale GA4GH (*Global alliance for genomics and health*) ayant

pour vocation la construction d'un système de partage des données pour faciliter l'obtention et l'interprétation clinique des données de santé. L'un des premiers projets développés, en collaboration avec le *Human Variome Project*, est le *BRCA Challenge*, dont l'objectif est de faire progresser la compréhension des bases moléculaires du syndrome seins-ovaires en mettant en commun l'ensemble des données sur les variations des gènes *BRCA1* et *BRCA2* ainsi que les données cliniques, et ce, à l'échelle mondiale (Cline *et al.*, 2018).

Avec la mise en place du Plan « Médecine France Génomique 2025 », la génomique est plus que jamais placée au cœur de l'innovation diagnostique, pronostique, thérapeutique et préventive pour une médecine de plus en plus personnalisée. Nul doute que cette nouvelle ère de la génomique en France va mettre exergue la nécessité de voir émerger des centres de diagnostics experts pour une pathologie donnée détenant un véritable savoir-faire dans l'interprétation des variations génétiques associées à la pathologie étudiée.

Partie V : Bibliographie

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Acedo, A., Sanz, D.J., Durán, M., Infante, M., Pérez-Cabornero, L., Miner, C., and Velasco, E.A. (2012). Comprehensive splicing functional analysis of DNA variants of the BRCA2 gene by hybrid minigenes. Breast Cancer Res. BCR *14*, R87.

Acedo, A., Hernández-Moro, C., Curiel-García, Á., Díez-Gómez, B., and Velasco, E.A. (2015). Functional classification of BRCA2 DNA variants by splicing assays in a large minigene with 9 exons. Hum. Mutat. *36*, 210–221.

Aebi, M., Hornig, H., and Weissmann, C. (1987). 5' cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5' splice region, not by the conserved 5' GU. Cell *50*, 237–246.

Aguilera, A., and García-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Mol. Cell *46*, 115–124.

Akerman, M., Fregoso, O.I., Das, S., Ruse, C., Jensen, M.A., Pappin, D.J., Zhang, M.Q., and Krainer, A.R. (2015). Differential connectivity of splicing activators and repressors to the human spliceosome. Genome Biol. *16*, 119.

Amendola, L.M., Jarvik, G.P., Leo, M.C., McLaughlin, H.M., Akkari, Y., Amaral, M.D., Berg, J.S., Biswas, S., Bowling, K.M., Conlin, L.K., *et al.* (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am. J. Hum. Genet. *99*, 247.

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D.A., Ghosh, S., and Jacobson, A. (2004). A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. Nature *432*, 112–118.

Anna, A., and Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. J. Appl. Genet. *59*, 253–268.

Antoniou, A.C., Foulkes, W.D., and Tischkowitz, M. (2014). Breast-cancer risk in families with mutations in PALB2. N. Engl. J. Med. *371*, 1651–1652.

Aparicio, T., Baer, R., and Gautier, J. (2014). DNA double-strand break repair pathway choice and cancer. DNA Repair *19*, 169–175.

Argente, J., Flores, R., Gutiérrez-Arumí, A., Verma, B., Martos-Moreno, G.Á., Cuscó, I., Oghabian, A., Chowen, J.A., Frilander, M.J., and Pérez-Jurado, L.A. (2014). Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. EMBO Mol. Med. *6*, 299–306.

Arnold, S., Buchanan, D.D., Barker, M., Jaskowski, L., Walsh, M.D., Birney, G., Woods, M.O., Hopper, J.L., Jenkins, M.A., Brown, M.A., *et al.* (2009). Classifying MLH1 and MSH2 variants

using bioinformatic prediction, splicing assays, segregation, and tumor characteristics. Hum. Mutat. *30*, 757–770.

Ars, E., Serra, E., García, J., Kruyer, H., Gaona, A., Lázaro, C., and Estivill, X. (2000). Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum. Mol. Genet. *9*, 237–247.

Badie, S., Escandell, J.M., Bouwman, P., Carlos, A.R., Thanasoula, M., Gallardo, M.M., Suram, A., Jaco, I., Benitez, J., Herbig, U., *et al.* (2010). BRCA2 acts as a RAD51 loader to facilitate telomere replication and capping. Nat. Struct. Mol. Biol. *17*, 1461–1469.

Bakker, J.L., Thirthagiri, E., van Mil, S.E., Adank, M.A., Ikeda, H., Verheul, H.M.W., Meijers-Heijboer, H., de Winter, J.P., Sharan, S.K., and Waisfisz, Q. (2014). A novel splice site mutation in the noncoding region of BRCA2: implications for Fanconi anemia and familial breast cancer diagnostics. Hum. Mutat. *35*, 442–446.

Baldeyron, C., Jacquemin, E., Smith, J., Jacquemont, C., De Oliveira, I., Gad, S., Feunteun, J., Stoppa-Lyonnet, D., and Papadopoulo, D. (2002). A single mutated BRCA1 allele leads to impaired fidelity of double strand break end-joining. Oncogene *21*, 1401–1410.

Baralle, D., and Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. J. Med. Genet. *42*, 737–748.

Baralle, D., and Buratti, E. (2017). RNA splicing in human disease and in the clinic. Clin. Sci. Lond. Engl. 1979 *131*, 355–368.

Baralle, D., Lucassen, A., and Buratti, E. (2009a). Missed threads. The impact of pre-mRNA splicing defects on clinical practice. EMBO Rep. *10*, 810–816.

Baralle, D., Lucassen, A., and Buratti, E. (2009b). Missed threads. The impact of pre-mRNA splicing defects on clinical practice. EMBO Rep. *10*, 810–816.

Baralle, M., Baralle, D., De Conti, L., Mattocks, C., Whittaker, J., Knezevich, A., Ffrench-Constant, C., and Baralle, F.E. (2003). Identification of a mutation that perturbs NF1 agene splicing using genomic DNA samples and a minigene assay. J. Med. Genet. *40*, 220–222.

Baralle, M., Skoko, N., Knezevich, A., De Conti, L., Motti, D., Bhuvanagiri, M., Baralle, D., Buratti, E., and Baralle, F.E. (2006). NF1 mRNA biogenesis: effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. FEBS Lett. *580*, 4449–4456.

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

Basile, D., Garattini, S.K., Bonotto, M., Ongaro, E., Casagrande, M., Cattaneo, M., Fanotto, V., De Carlo, E., Loupakis, F., Urbano, F., *et al.* (2017). Immunotherapy for colorectal cancer: where are we heading? Expert Opin. Biol. Ther. *17*, 709–721.

Belgrader, P., Cheng, J., and Maquat, L.E. (1993). Evidence to implicate translation by ribosomes in the mechanism by which nonsense codons reduce the nuclear level of human triosephosphate isomerase mRNA. Proc. Natl. Acad. Sci. U. S. A. *90*, 482–486.

Bennett, C.B., Westmoreland, T.J., Snipe, J.R., and Resnick, M.A. (1996). A double-strand break within a yeast artificial chromosome (YAC) containing human DNA can result in YAC loss, deletion or cell lethality. Mol. Cell. Biol. *16*, 4414–4425.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. *270*, 2411–2414.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. U. S. A. *74*, 3171–3175.

Bever, K.M., and Le, D.T. (2017). An Expanding Role for Immunotherapy in Colorectal Cancer. J. Natl. Compr. Cancer Netw. JNCCN *15*, 401–410.

Bianchi, F., Raponi, M., Piva, F., Viel, A., Bearzi, I., Galizia, E., Bracci, R., Belvederesi, L., Loretelli, C., Brugiati, C., *et al.* (2011). An intronic mutation in MLH1 associated with familial colon and breast cancer. Fam. Cancer *10*, 27–35.

Binder-Foucard, F., Bossard, N., Delafosse, P., Belot, A., Woronoff, A.-S., Remontet, L., and French network of cancer registries (Francim) (2014). Cancer incidence and mortality in France over the 1980-2012 period: solid tumors. Rev. Epidemiol. Sante Publique *62*, 95–108.

Biswas, K., Das, R., Alter, B.P., Kuznetsov, S.G., Stauffer, S., North, S.L., Burkett, S., Brody, L.C., Meyer, S., Byrd, R.A., *et al.* (2011). A comprehensive functional characterization of BRCA2 variants associated with Fanconi anemia using mouse ES cell-based assay. Blood *118*, 2430–2442.

Biswas, K., Das, R., Eggington, J.M., Qiao, H., North, S.L., Stauffer, S., Burkett, S.S., Martin, B.K., Southon, E., Sizemore, S.C., *et al.* (2012). Functional evaluation of BRCA2 variants mapping to the PALB2-binding and C-terminal DNA-binding domains using a mouse ES cell-based assay. Hum. Mol. Genet. *21*, 3993–4006.

Bitton, D.A., Rallis, C., Jeffares, D.C., Smith, G.C., Chen, Y.Y.C., Codlin, S., Marguerat, S., and Bähler, J. (2014). LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. Genome Res. *24*, 1169–1179.

Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell *126*, 37–47.

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., *et al.* (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096.

Boland, C.R., and Troncale, F.J. (1984). Familial colonic cancer without antecedent polyposis. Ann. Intern. Med. *100*, 700–701.

Boland, C.R., Thibodeau, S.N., Hamilton, S.R., Sidransky, D., Eshleman, J.R., Burt, R.W., Meltzer, S.J., Rodriguez-Bigas, M.A., Fodde, R., Ranzani, G.N., *et al.* (1998). A National Cancer

Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. *58*, 5248–5257.

Bonadona, V., Bonaïti, B., Olschwang, S., Grandjouan, S., Huiart, L., Longy, M., Guimbaud, R., Buecher, B., Bignon, Y.-J., Caron, O., *et al.* (2011). Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. JAMA *305*, 2304–2310.

Bonatti, F., Pepe, C., Tancredi, M., Lombardi, G., Aretini, P., Sensi, E., Falaschi, E., Cipollini, G., Bevilacqua, G., and Caligo, M.A. (2006). RNA-based analysis of BRCA1 and BRCA2 gene alterations. Cancer Genet. Cytogenet. *170*, 93–101.

Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., Berthet, P., Chevrier, A., Dugast, C., Layet, V., *et al.* (2008). 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. *45*, 438–446.

Bourgeois, C.F., Lejeune, F., and Stévenin, J. (2004). Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. Prog. Nucleic Acid Res. Mol. Biol. *78*, 37–88.

Bousquet-Antonelli, C., Presutti, C., and Tollervey, D. (2000). Identification of a regulated pathway for nuclear pre-mRNA turnover. Cell *102*, 765–775.

Bouwman, P., van der Gulden, H., van der Heijden, I., Drost, R., Klijn, C.N., Prasetyanti, P., Pieterse, M., Wientjens, E., Seibler, J., Hogervorst, F.B.L., *et al.* (2013). A high-throughput functional complementation assay for classification of BRCA1 missense variants. Cancer Discov. *3*, 1142–1155.

Brandão, R.D., Mensaert, K., López-Perolio, I., Tserpelis, D., Xenakis, M., Lattimore, V., Walker, L.C., Kvist, A., Vega, A., Gutiérrez-Enríquez, S., *et al.* (2019). Targeted RNA-seq successfully identifies normal and pathogenic splicing events in breast/ovarian cancer susceptibility and Lynch syndrome genes. Int. J. Cancer.

Brogna, S., McLeod, T., and Petric, M. (2016). The Meaning of NMD: Translate or Perish. Trends Genet. TIG *32*, 395–407.

Brookes, A.J., and Robinson, P.N. (2015). Human genotype-phenotype databases: aims, challenges and opportunities. Nat. Rev. Genet. *16*, 702–715.

Bubien, V., Bonnet, F., Brouste, V., Hoppe, S., Barouk-Simonet, E., David, A., Edery, P., Bottani, A., Layet, V., Caron, O., *et al.* (2013). High cumulative risks of cancer in patients with PTEN hamartoma tumour syndrome. J. Med. Genet. *50*, 255–263.

Buchanan, D.D., Tan, Y.Y., Walsh, M.D., Clendenning, M., Metcalf, A.M., Ferguson, K., Arnold, S.T., Thompson, B.A., Lose, F.A., Parsons, M.T., *et al.* (2014). Tumor Mismatch Repair Immunohistochemistry and DNA MLH1 Methylation Testing of Patients With Endometrial Cancer

Diagnosed at Age Younger Than 60 Years Optimizes Triage for Population-Level Germline Mismatch Repair Gene Mutation Testing. J. Clin. Oncol. *32*, 90–100.

Buckland, P.R. (2004). Allele-specific gene expression differences in humans. Hum. Mol. Genet. *13 Spec No 2*, R255-260.

Buecher, B., and Laurent-Puig, P. (2010). Les formes héréditaires non polyposiques des cancers colorectaux. *17*, 8.

Buecher, B., de, P., Antoine, Freneaux, P., and Rouleau, E. (2011). Instabilité des microsatellites et cancers colorectaux. Cancéro Dig.

Byrjalsen, A., Steffensen, A.Y., Hansen, T.V.O., Wadt, K., and Gerdes, A.-M. (2017). 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. *5*, 876–879.

Byrnes, G.B., Southey, M.C., and Hopper, J.L. (2008). Are the so-called low penetrance breast cancer genes, ATM, BRIP1, PALB2 and CHEK2, high risk for women with strong family histories? Breast Cancer Res. BCR *10*, 208.

Caminsky, N., Mucaki, E.J., and Rogan, P.K. (2014). Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Research *3*, 282.

Caputo, S., Benboudjema, L., Sinilnikova, O., Rouleau, E., Béroud, C., Lidereau, R., and French BRCA GGC Consortium (2012). Description and analysis of genetic variants in French hereditary breast and ovarian cancer families recorded in the UMD-BRCA1/BRCA2 databases. Nucleic Acids Res. *40*, D992-1002.

Caputo, S.M., Léone, M., Damiola, F., Ehlen, A., Carreira, A., Gaidrat, P., Martins, A., Brandão, R.D., Peixoto, A., Vega, A., *et al.* (2018). Full in-frame exon 3 skipping of BRCA2 confers high risk of breast and/or ovarian cancer. Oncotarget *9*, 17334–17348.

Carethers, J.M., Chauhan, D.P., Fink, D., Nebel, S., Bresalier, R.S., Howell, S.B., and Boland, C.R. (1999). Mismatch repair proficiency and in vitro response to 5-fluorouracil. Gastroenterology *117*, 123–131.

Cartegni, L., and Krainer, A.R. (2002). Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nat. Genet. *30*, 377–384.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. *31*, 3568–3571.

Cartegni, L., Hastings, M.L., Calarco, J.A., de Stanchina, E., and Krainer, A.R. (2006). Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. Am. J. Hum. Genet. *78*, 63–77.

Carter, M.S., Doskow, J., Morris, P., Li, S., Nhim, R.P., Sandstedt, S., and Wilkinson, M.F. (1995). A regulatory mechanism that detects premature nonsense codons in T-cell receptor transcripts in vivo is reversed by protein synthesis inhibitors in vitro. J. Biol. Chem. *270*, 28995–29003.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. *40*, 1416–1425.

Caux-Moncoutier, V., Pagès-Berhouet, S., Michaux, D., Asselain, B., Castéra, L., De Pauw, A., Buecher, B., Gauthier-Villars, M., Stoppa-Lyonnet, D., and Houdayer, C. (2009). Impact of BRCA1 and BRCA2 variants on splicing: clues from an allelic imbalance study. Eur. J. Hum. Genet. EJHG *17*, 1471–1480.

Chang, S., Biswas, K., Martin, B.K., Stauffer, S., and Sharan, S.K. (2009). Expression of human BRCA1 variants in mouse ES cells allows functional analysis of BRCA1 mutations. J. Clin. Invest. *119*, 3160–3171.

Chang, Y.-F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. *76*, 51–74.

Charbonnier, F., Martin, C., Scotte, M., Sibert, L., Moreau, V., and Frebourg, T. (1995). Alternative splicing of MLH1 messenger RNA in human normal cells. Cancer Res. *55*, 1839–1841.

Chasin, L.A. (2007). Searching for splicing motifs. Adv. Exp. Med. Biol. *623*, 85–106.

Chen, H.-C., and Cheng, S.-C. (2012). Functional roles of protein splicing factors. Biosci. Rep. *32*, 345–359.

Chen, S., and Parmigiani, G. (2007). Meta-Analysis of BRCA1 and BRCA2 Penetrance. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *25*, 1329–1333.

Cheng, J., Belgrader, P., Zhou, X., and Maquat, L.E. (1994). Introns are cis effectors of the nonsense-codon-mediated reduction in nuclear mRNA abundance. Mol. Cell. Biol. *14*, 6317–6325.

Chiang, H.-L., Wu, J.-Y., and Chen, Y.-T. (2017). Identification of functional single nucleotide polymorphisms in the branchpoint site. Hum. Genomics *11*, 27.

Chillón, M., Dörk, T., Casals, T., Giménez, J., Fonknechten, N., Will, K., Ramos, D., Nunes, V., and Estivill, X. (1995). A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA-->G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. Am. J. Hum. Genet. *56*, 623–629.

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

Claus, E.B., Risch, N., and Thompson, W.D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. Am. J. Hum. Genet. *48*, 232–242.

Clement, S.L., and Lykke-Andersen, J. (2006). No mercy for messages that mess with the ribosome. Nat. Struct. Mol. Biol. *13*, 299–301.

Cline, M.S., Liao, R.G., Parsons, M.T., Paten, B., Alquaddoomi, F., Antoniou, A., Baxter, S., Brody, L., Cook-Deegan, R., Coffin, A., *et al.* (2018). BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. PLoS Genet. *14*, e1007752.

Colombo, M., Blok, M.J., Whiley, P., Santamariña, M., Gutiérrez-Enríquez, S., Romero, A., Garre, P., Becker, A., Smith, L.D., De Vecchi, G., *et al.* (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. Hum. Mol. Genet. *23*, 3666–3680.

Colombo, M., Lòpez-Perolio, I., Meeks, H.D., Caleca, L., Parsons, M.T., Li, H., De Vecchi, G., Tudini, E., Foglia, C., Mondini, P., *et al.* (2018). The BRCA2 c.68-7T > A variant is not pathogenic: A model for clinical calibration of spliceogenicity. Hum. Mutat. *39*, 729–741.

Cooper, T.A. (2005). Use of minigene systems to dissect alternative splicing elements. Methods San Diego Calif *37*, 331–340.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Corvelo, A., Hallegger, M., Smith, C.W.J., and Eyras, E. (2010). Genome-wide association between branch point properties and alternative splicing. PLoS Comput. Biol. *6*, e1001016.

da Costa, P.J., Menezes, J., and Romão, L. (2017). The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. Int. J. Biochem. Cell Biol. *91*, 168–175.

Cousineau, I., and Belmaaza, A. (2007). BRCA1 haploinsufficiency, but not heterozygosity for a BRCA1-truncating mutation, deregulates homologous recombination. Cell Cycle Georget. Tex *6*, 962–971.

Culbertson, M.R., and Leeds, P.F. (2003). Looking at mRNA decay pathways through the window of molecular evolution. Curr. Opin. Genet. Dev. *13*, 207–214.

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

David, C.J., Chen, M., Assanah, M., Canoll, P., and Manley, J.L. (2010). HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature *463*, 364–368.

Davy, G., Rousselin, A., Goardon, N., Castéra, L., Harter, V., Legros, A., Muller, E., Fouillet, R., Brault, B., Smirnova, A.S., *et al.* (2017). Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. Eur. J. Hum. Genet. EJHG *25*, 1147–1154.

De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA *4*, 49–60.

Deans, A.J., and West, S.C. (2011). DNA interstrand crosslink repair and cancer. Nat. Rev. Cancer *11*, 467–480.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.

Dhir, A., and Buratti, E. (2010). Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. FEBS J. *277*, 841–855.

Di Blasi, C., He, Y., Morandi, L., Cornelio, F., Guicheney, P., and Mora, M. (2001). Mild muscular dystrophy due to a nonsense mutation in the LAMA2 gene resulting in exon skipping. Brain J. Neurol. *124*, 698–704.

Di Giacomo, D., Gaildrat, P., Abuli, A., Abdat, J., Frébourg, T., Tosi, M., and Martins, A. (2013). 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. *34*, 1547–1557.

Dimaano, C., and Ullman, K.S. (2004). Nucleocytoplasmic transport: integrating mRNA production and turnover with export through the nuclear pore. Mol. Cell. Biol. *24*, 3069–3076.

Disset, A., Bourgeois, C.F., Benmalek, N., Claustres, M., Stevenin, J., and Tuffery-Giraud, S. (2006). An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. Hum. Mol. Genet. *15*, 999–1013.

Dorman, S.N., Viner, C., and Rogan, P.K. (2014). Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. Sci. Rep. *4*, 7063.

Dujardin, G., Daguenet, É., Bernard, D.G., Flodrops, M., Durand, S., Chauveau, A., El Khoury, F., Le Jossic-Corcos, C., and Corcos, L. (2016). [Pre-mRNA splicing: when the spliceosome loses ground]. Med. Sci. MS *32*, 1103–1110.

Durand, S., Cougot, N., Mahuteau-Betzer, F., Nguyen, C.-H., Grierson, D.S., Bertrand, E., Tazi, J., and Lejeune, F. (2007). Inhibition of nonsense-mediated mRNA decay (NMD) by a new chemical molecule reveals the dynamic of NMD factors in P-bodies. J. Cell Biol. *178*, 1145–1160.

Durkacz, B.W., Omidiji, O., Gray, D.A., and Shall, S. (1980). (ADP-ribose)n participates in DNA excision repair. Nature *283*, 593–596.

ElShamy, W.M., and Livingston, D.M. (2004). Identification of BRCA1-IRIS, a BRCA1 locus product. Nat. Cell Biol. *6*, 954–967.

Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J.O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res. *42*, 10681–10697.

Fabbro, M., Rodriguez, J.A., Baer, R., and Henderson, B.R. (2002). BARD1 induces BRCA1 intranuclear foci formation by increasing RING-dependent BRCA1 nuclear import and inhibiting BRCA1 nuclear export. J. Biol. Chem. *277*, 21315–21324.

Fackenthal, J.D., Yoshimatsu, T., Zhang, B., de Garibay, G.R., Colombo, M., De Vecchi, G., Ayoub, S.C., Lal, K., Olopade, O.I., Vega, A., *et al*. (2016). Naturally occurring BRCA2 alternative mRNA splicing events in clinically relevant samples. J. Med. Genet. *53*, 548–558.

Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007–1013.

Fairbrother, W.G., Holste, D., Burge, C.B., and Sharp, P.A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. *2*, E268.

Fan, S., Yuan, R., Ma, Y.X., Meng, Q., Goldberg, I.D., and Rosen, E.M. (2001). Mutant BRCA1 genes antagonize phenotype of wild-type BRCA1. Oncogene *20*, 8215–8235.

Favoriti, P., Carbone, G., Greco, M., Pirozzi, F., Pirozzi, R.E.M., and Corcione, F. (2016). Worldwide burden of colorectal cancer: a review. Updat. Surg. *68*, 7–11.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.

Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature.

Fishel, R., Lescoe, M.K., Rao, M.R., Copeland, N.G., Jenkins, N.A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell *75*, 1027–1038.

Flanigan, K.M., Dunn, D.M., von Niederhausern, A., Soltanzadeh, P., Howard, M.T., Sampson, J.B., Swoboda, K.J., Bromberg, M.B., Mendell, J.R., Taylor, L.E., *et al.* (2011). Nonsense mutation-associated Becker muscular dystrophy: interplay between exon definition and splicing regulatory elements within the DMD gene. Hum. Mutat. *32*, 299–308.

Flouriot, G., Brand, H., Seraphin, B., and Gannon, F. (2002). Natural trans-spliced mRNAs are generated from the human estrogen receptor-alpha (hER alpha) gene. J. Biol. Chem. *277*, 26244–26251.

Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.-P., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. U. S. A. *102*, 16176–16181.

Fradet-Turcotte, A., Sitz, J., Grapton, D., and Orthwein, A. (2016). BRCA2 functions: from DNA repair to replication fork stabilization. Endocr. Relat. Cancer *23*, T1–T17.

Fraile-Bethencourt, E., Díez-Gómez, B., Velásquez-Zapata, V., Acedo, A., Sanz, D.J., and Velasco, E.A. (2017). Functional classification of DNA variants by hybrid minigenes: Identification of 30 spliceogenic variants of BRCA2 exons 17 and 18. PLoS Genet. *13*, e1006691.

Frebourg, T. (2014). The challenge for the next generation of medical geneticists. Hum. Mutat. *35*, 909–911.

Fredericks, A.M., Cygan, K.J., Brown, B.A., and Fairbrother, W.G. (2015). RNA-Binding Proteins: Splicing Factors and Disease. Biomolecules *5*, 893–909.

Friedberg, E.C. (2003). DNA damage and repair. Nature *421*, 436–440.

Frischmeyer, P.A., van Hoof, A., O'Donnell, K., Guerrerio, A.L., Parker, R., and Dietz, H.C. (2002). An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science *295*, 2258–2261.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. Clifton NJ *653*, 249–257.

Gaildrat, P., Krieger, S., Di Giacomo, D., Abdat, J., Révillion, F., Caputo, S., Vaur, D., Jamard, E., Bohers, E., Ledemeney, D., *et al.* (2012). Multiple sequence variants of BRCA2 exon 7 alter splicing regulation. J. Med. Genet. *49*, 609–617.

Gallie, D.R. (1998). A tale of two termini: a functional interaction between the termini of an mRNA is a prerequisite for efficient translation initiation. Gene *216*, 1–11.

Ganesan, S., Silver, D.P., Greenberg, R.A., Avni, D., Drapkin, R., Miron, A., Mok, S.C., Randrianarison, V., Brodie, S., Salstrom, J., *et al.* (2002). BRCA1 supports XIST RNA concentration on the inactive X chromosome. Cell *111*, 393–405.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res. *36*, 2257–2267.

Garber, J.E., and Offit, K. (2005). Hereditary cancer predisposition syndromes. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *23*, 276–292.

Garcia-Higuera, I., Taniguchi, T., Ganesan, S., Meyn, M.S., Timmers, C., Hejna, J., Grompe, M., and D'Andrea, A.D. (2001). Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. Mol. Cell *7*, 249–262.

Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nat. Protoc. *11*, 1782–1787.

Gatfield, D., Unterholzner, L., Ciccarelli, F.D., Bork, P., and Izaurralde, E. (2003). Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. EMBO J. *22*, 3960–3970.

Gehring, N.H., Neu-Yilik, G., Schell, T., Hentze, M.W., and Kulozik, A.E. (2003). Y14 and hUpf3b form an NMD-activating complex. Mol. Cell *11*, 939–949.

Geib, T., and Hertel, K.J. (2009). Restoration of full-length SMN promoted by adenoviral vectors expressing RNA antisense oligonucleotides embedded in U7 snRNAs. PloS One *4*, e8204.

Genuardi, M., Viel, A., Bonora, D., Capozzi, E., Bellacosa, A., Leonardi, F., Valle, R., Ventura, A., Pedroni, M., Boiocchi, M., *et al.* (1998). Characterization of MLH1 and MSH2 alternative splicing and its relevance to molecular testing of colorectal cancer susceptibility. Hum. Genet. *102*, 15–20.

Gerbino, V., Carrì, M.T., Cozzolino, M., and Achsel, T. (2013). Mislocalised FUS mutants stall spliceosomal snRNPs in the cytoplasm. Neurobiol. Dis. *55*, 120–128.

Geuens, T., Bouhy, D., and Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. Hum. Genet. *135*, 851–867.

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P.M., Green, M.R., Riva, S., and Biamonti, G. (2005). Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. Mol. Cell *20*, 881–890.

Giardiello, F.M., Brensinger, J.D., Tersmette, A.C., Goodman, S.N., Petersen, G.M., Booker, S.V., Cruz-Correa, M., and Offerhaus, J.A. (2000). Very high risk of cancer in familial Peutz-Jeghers syndrome. Gastroenterology *119*, 1447–1453.

Gilbert, W. (1978a). Why genes in pieces? Nature *271*, 501.

Gilbert, W. (1978b). Why genes in pieces? Nature *271*, 501.

Girard, C., Will, C.L., Peng, J., Makarov, E.M., Kastner, B., Lemm, I., Urlaub, H., Hartmuth, K., and Lührmann, R. (2012). Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. Nat. Commun. *3*, 994.

Glickman, B.W., and Radman, M. (1980). Escherichia coli mutator mutants deficient in methylation-instructed DNA mismatch correction. Proc. Natl. Acad. Sci. U. S. A. *77*, 1063–1067.

Goina, E., Skoko, N., and Pagani, F. (2008). Binding of DAZAP1 and hnRNPA1/A2 to an exonic splicing silencer in a natural BRCA1 exon 18 mutant. Mol. Cell. Biol. *28*, 3850–3860.

Goldgar, D.E., Easton, D.F., Deffenbaugh, A.M., Monteiro, A.N.A., Tavtigian, S.V., Couch, F.J., and Breast Cancer Information Core (BIC) Steering Committee (2004). Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am. J. Hum. Genet. *75*, 535–544.

Golmard, L., Caux-Moncoutier, V., Davy, G., Al Ageeli, E., Poirot, B., Tirapo, C., Michaux, D., Barbaroux, C., d'Enghien, C.D., Nicolas, A., *et al.* (2013). Germline mutation in the RAD51B gene confers predisposition to breast cancer. BMC Cancer *13*, 484.

Gong, Z., Kim, J.-E., Leung, C.C.Y., Glover, J.N.M., and Chen, J. (2010). BACH1/FANCJ acts with TopBP1 and participates early in DNA replication checkpoint control. Mol. Cell *37*, 438–446.

González, C.I., Ruiz-Echevarría, M.J., Vasudevan, S., Henry, M.F., and Peltz, S.W. (2000). The yeast hnRNP-like protein Hrp1/Nab4 marks a transcript for nonsense-mediated mRNA decay. Mol. Cell *5*, 489–499.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. Mol. Cell *22*, 769–781.

Gorman, L., Suter, D., Emerick, V., Schümperli, D., and Kole, R. (1998). Stable alteration of pre-mRNA splicing patterns by modified U7 small nuclear RNAs. Proc. Natl. Acad. Sci. U. S. A. *95*, 4929–4934.

Goyenvalle, A., Vulin, A., Fougerousse, F., Leturcq, F., Kaplan, J.-C., Garcia, L., and Danos, O. (2004). Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping. Science *306*, 1796–1799.

Grandval, P., Fabre, A.J., Gaildrat, P., Baert-Desurmont, S., Buisine, M.-P., Ferrari, A., Wang, Q., Béroud, C., and Olschwang, S. (2013). UMD-MLH1/MSH2/MSH6 databases: description and analysis of genetic variations in French Lynch syndrome families. Database J. Biol. Databases Curation *2013*, bat036.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA N. Y. N *6*, 1197–1211.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. TIG *17*, 100–107.

Guidugli, L., Carreira, A., Caputo, S.M., Ehlen, A., Galli, A., Monteiro, A.N.A., Neuhausen, S.L., Hansen, T.V.O., Couch, F.J., Vreeswijk, M.P.G., *et al.* (2014). Functional assays for analysis of variants of uncertain significance in BRCA2. Hum. Mutat. *35*, 151–164.

Guidugli, L., Shimelis, H., Masica, D.L., Pankratz, V.S., Lipton, G.B., Singh, N., Hu, C., Monteiro, A.N.A., Lindor, N.M., Goldgar, D.E., *et al.* (2018). Assessment of the Clinical Relevance of BRCA2 Missense Variants by Functional and Computational Approaches. Am. J. Hum. Genet.

Guillotin, D., and Martin, S.A. (2014). Exploiting DNA mismatch repair deficiency as a therapeutic strategy. Exp. Cell Res. *329*, 110–115.

Hakem, R., de la Pompa, J.L., Sirard, C., Mo, R., Woo, M., Hakem, A., Wakeham, A., Potter, J., Reitmair, A., Billia, F., *et al.* (1996). The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. Cell *85*, 1009–1023.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., and King, M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science *250*, 1684–1689.

Hammond, S.M., and Wood, M.J.A. (2011). Genetic therapies for RNA mis-splicing diseases. Trends Genet. TIG *27*, 196–205.

Han, L., Vickers, K.C., Samuels, D.C., and Guo, Y. (2015). Alternative applications for distinct RNA sequencing strategies. Brief. Bioinform. *16*, 629–639.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Harigaya, Y., and Parker, R. (2010). No-go decay: a quality control mechanism for RNA in translation. Wiley Interdiscip. Rev. RNA *1*, 132–141.

Hartford, S.A., Chittela, R., Ding, X., Vyas, A., Martin, B., Burkett, S., Haines, D.C., Southon, E., Tessarollo, L., and Sharan, S.K. (2016). Interaction with PALB2 Is Essential for Maintenance of Genomic Integrity by BRCA2. PLoS Genet. *12*, e1006236.

Hartmann, L., Theiss, S., Niederacher, D., and Schaal, H. (2008). Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases? Front. Biosci. J. Virtual Libr. *13*, 3252–3272.

Hastings, M.L., Berniac, J., Liu, Y.H., Abato, P., Jodelka, F.M., Barthel, L., Kumar, S., Dudley, C., Nelson, M., Larson, K., *et al.* (2009). Tetracyclines that promote SMN2 exon 7 splicing as therapeutics for spinal muscular atrophy. Sci. Transl. Med. *1*, 5ra12.

Havens, M.A., Duelli, D.M., and Hastings, M.L. (2013). Targeting RNA splicing for disease therapy. Wiley Interdiscip. Rev. RNA *4*, 247–266.

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

Hendriks, G., Morolli, B., Calléja, F.M.G.R., Plomp, A., Mesman, R.L.S., Meijers, M., Sharan, S.K., Vreeswijk, M.P.G., and Vrieling, H. (2014). An efficient pipeline for the generation and functional analysis of human BRCA2 variants of uncertain significance. Hum. Mutat. *35*, 1382–1391.

Hendriks, Y.M.C., Wagner, A., Morreau, H., Menko, F., Stormorken, A., Quehenberger, F., Sandkuijl, L., Møller, P., Genuardi, M., Van Houwelingen, H., *et al.* (2004). Cancer risk in hereditary nonpolyposis colorectal cancer due to MSH6 mutations: impact on counseling and surveillance. Gastroenterology *127*, 17–25.

Hernández-Imaz, E., Martín, Y., de Conti, L., Melean, G., Valero, A., Baralle, M., and Hernández-Chico, C. (2015). Functional Analysis of Mutations in Exon 9 of NF1 Reveals the Presence of Several Elements Regulating Splicing. PloS One *10*, e0141735.

Hilbers, F.S., Wijnen, J.T., Hoogerbrugge, N., Oosterwijk, J.C., Collee, M.J., Peterlongo, P., Radice, P., Manoukian, S., Feroce, I., Capra, F., *et al.* (2012). Rare variants in XRCC2 as breast cancer susceptibility alleles. J. Med. Genet. *49*, 618–620.

Hilleren, P., McCarthy, T., Rosbash, M., Parker, R., and Jensen, T.H. (2001). Quality control of mRNA 3'-end processing is linked to the nuclear exosome. Nature *413*, 538–542.

Hinzpeter, A., Aissat, A., Sondo, E., Costa, C., Arous, N., Gameiro, C., Martin, N., Tarze, A., Weiss, L., de Becdelièvre, A., *et al.* (2010). Alternative splicing at a NAGNAG acceptor site as a novel phenotype modifier. PLoS Genet. *6*.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., and Pease, L.R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene *77*, 51–59.

Hoffman, J.D., Hallam, S.E., Venne, V.L., Lyon, E., and Ward, K. (1998). Implications of a novel cryptic splice site in the BRCA1 gene. Am. J. Med. Genet. *80*, 140–144.

van Hoof, A., Frischmeyer, P.A., Dietz, H.C., and Parker, R. (2002). Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science *295*, 2262–2264.

Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., *et al.* (2012). 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. *33*, 1228–1238.

Howard, J.M., and Sanford, J.R. (2015). The RNAissance family: SR proteins as multifaceted regulators of gene expression. Wiley Interdiscip. Rev. RNA *6*, 93–110.

Howlett, N.G., Taniguchi, T., Olson, S., Cox, B., Waisfisz, Q., De Die-Smulders, C., Persky, N., Grompe, M., Joenje, H., Pals, G., *et al.* (2002). Biallelic inactivation of BRCA2 in Fanconi anemia. Science *297*, 606–609.

de la Hoya, M., Soukarieh, O., López-Perolio, I., Vega, A., Walker, L.C., van Ierland, Y., Baralle, D., Santamariña, M., Lattimore, V., Wijnen, J., *et al.* (2016a). 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. *25*, 2256–2268.

de la Hoya, M., Soukarieh, O., López-Perolio, I., Vega, A., Walker, L.C., van Ierland, Y., Baralle, D., Santamariña, M., Lattimore, V., Wijnen, J., *et al.* (2016b). 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. *25*, 2256–2268.

Hsieh, P., and Yamane, K. (2008). DNA mismatch repair: molecular mechanism, cancer, and ageing. Mech. Ageing Dev. *129*, 391–407.

Hui, J., Hung, L.-H., Heiner, M., Schreiner, S., Neumüller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. EMBO J. *24*, 1988–1998.

Hung, L.-H., Heiner, M., Hui, J., Schreiner, S., Benes, V., and Bindereif, A. (2008). Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. RNA N. Y. N *14*, 284–296.

Hutton, M., Lendon, C.L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A., *et al.* (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. Nature *393*, 702–705.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Ishigaki, Y., Li, X., Serin, G., and Maquat, L.E. (2001). Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. Cell *106*, 607–617.

Isken, O., Kim, Y.K., Hosoda, N., Mayeur, G.L., Hershey, J.W.B., and Maquat, L.E. (2008). Upf1 phosphorylation triggers translational repression during nonsense-mediated mRNA decay. Cell *133*, 314–327.

Iyer, R.R., Pluciennik, A., Burdett, V., and Modrich, P.L. (2006). DNA mismatch repair: functions and mechanisms. Chem. Rev. *106*, 302–323.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. *19*, 3795–3798.

Jarvik, G.P., and Browning, B.L. (2016). Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants. Am. J. Hum. Genet. *98*, 1077–1081.

Jasin, M., and Rothstein, R. (2013). Repair of strand breaks by homologous recombination. Cold Spring Harb. Perspect. Biol. *5*, a012740.

Jeggo, P.A., and Löbrich, M. (2007). DNA double-strand breaks: their cellular and clinical impact? Oncogene *26*, 7717–7719.

Jenkins, R.N., Osborne-Lawrence, S.L., Sinclair, A.K., Eddy, R.L., Byers, M.G., Shows, T.B., and Duby, A.D. (1990). Structure and chromosomal location of the human gene encoding cartilage matrix protein. J. Biol. Chem. *265*, 19624–19631.

Jensen, C.J., Oldfield, B.J., and Rubio, J.P. (2009). Splicing, cis genetic variation and disease. Biochem. Soc. Trans. *37*, 1311–1315.

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. Mol. Cells *40*, 1–9.

Jiricny, J. (2013). Postreplicative mismatch repair. Cold Spring Harb. Perspect. Biol. *5*, a012633.

Johnston, J.J., and Biesecker, L.G. (2013). Databases of genomic variation and phenotypes: existing resources and future needs. Hum. Mol. Genet. *22*, R27-31.

Jover, R., Zapater, P., Castells, A., Llor, X., Andreu, M., Cubiella, J., Balaguer, F., Sempere, L., Xicola, R.M., Bujanda, L., *et al.* (2009). The efficacy of adjuvant chemotherapy with 5-fluorouracil in colorectal cancer depends on the mismatch repair status. Eur. J. Cancer Oxf. Engl. 1990 *45*, 365–373.

Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat. Commun. *7*, 11558.

Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., and Krainer, A.R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat. Struct. Mol. Biol. *14*, 185–193.

Karousis, E.D., Nasif, S., and Mühlemann, O. (2016). Nonsense-mediated mRNA decay: novel mechanistic insights and biological impact. Wiley Interdiscip. Rev. RNA *7*, 661–682.

Kashima, T., and Manley, J.L. (2003). A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. Nat. Genet. *34*, 460–463.

Kashima, I., Yamashita, A., Izumi, N., Kataoka, N., Morishita, R., Hoshino, S., Ohno, M., Dreyfuss, G., and Ohno, S. (2006). Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. Genes Dev. *20*, 355–367.

Kashima, T., Rao, N., David, C.J., and Manley, J.L. (2007). hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. Hum. Mol. Genet. *16*, 3149–3159.

Kast, K., Rhiem, K., Wappenschmidt, B., Hahnen, E., Hauke, J., Bluemcke, B., Zarghooni, V., Herold, N., Ditsch, N., Kiechle, M., *et al.* (2016). Prevalence of BRCA1/2 germline mutations in 21 401 families with breast and ovarian cancer. J. Med. Genet. *53*, 465–471.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Ke, S., Anquetil, V., Zamalloa, J.R., Maity, A., Yang, A., Arias, M.A., Kalachikov, S., Russo, J.J., Ju, J., and Chasin, L.A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. *28*, 11–24.

Kee, Y., and D'Andrea, A.D. (2012). Molecular pathogenesis and clinical management of Fanconi anemia. J. Clin. Invest. *122*, 3799–3806.

Kerényi, Z., Mérai, Z., Hiripi, L., Benkovics, A., Gyula, P., Lacomme, C., Barta, E., Nagy, F., and Silhavy, D. (2008). Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. EMBO J. *27*, 1585–1595.

Kim, Y.-J., and Kim, H.-S. (2012). Alternative splicing and its impact as a cancer diagnostic marker. Genomics Inform. *10*, 74–80.

Kim, E., Ilagan, J.O., Liang, Y., Daubner, G.M., Lee, S.C.-W., Ramakrishnan, A., Li, Y., Chung, Y.R., Micol, J.-B., Murphy, M.E., *et al.* (2015). SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. Cancer Cell *27*, 617–630.

Kishore, S., Khanna, A., and Stamm, S. (2008). Rapid generation of splicing reporters with pSpliceExpress. Gene *427*, 104–110.

van der Klift, H.M., Jansen, A.M.L., van der Steenstraten, N., Bik, E.C., Tops, C.M.J., Devilee, P., and Wijnen, J.T. (2015). 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. *3*, 327–345.

Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U. S. A. *68*, 820–823.

Ko, M.J., Murata, K., Hwang, D.-S., and Parvin, J.D. (2006). Inhibition of BRCA1 in breast cell lines causes the centrosome duplication cycle to be disconnected from the cell cycle. Oncogene *25*, 298–303.

Kobayashi, H., Ohno, S., Sasaki, Y., and Matsuura, M. (2013). Hereditary breast and ovarian cancer susceptibility genes (review). Oncol. Rep. *30*, 1019–1029.

Komeno, Y., Huang, Y.-J., Qiu, J., Lin, L., Xu, Y., Zhou, Y., Chen, L., Monterroza, D.D., Li, H., DeKelver, R.C., *et al.* (2015). SRSF2 Is Essential for Hematopoiesis, and Its Myelodysplastic Syndrome-Related Mutations Dysregulate Alternative Pre-mRNA Splicing. Mol. Cell. Biol. *35*, 3071–3082.

Kong-Beltran, M., Seshagiri, S., Zha, J., Zhu, W., Bhawe, K., Mendoza, N., Holcomb, T., Pujara, K., Stinson, J., Fu, L., *et al.* (2006). Somatic mutations lead to an oncogenic deletion of met in lung cancer. Cancer Res. *66*, 283–289.

Konishi, H., Mohseni, M., Tamaki, A., Garay, J.P., Croessmann, S., Karnan, S., Ota, A., Wong, H.Y., Konishi, Y., Karakas, B., *et al.* (2011). Mutation of a single allele of the cancer susceptibility gene BRCA1 leads to genomic instability in human breast epithelial cells. Proc. Natl. Acad. Sci. U. S. A. *108*, 17773–17778.

Kowalewski, C., Bremer, J., Gostynski, A., Wertheim-Tysarowska, K., Wozniak, K., Bal, J., Jonkman, M.F., and Pasmooij, A.M.G. (2016). Amelioration of junctional epidermolysis bullosa due to exon skipping. Br. J. Dermatol. *174*, 1375–1379.

Královicová, J., Lei, H., and Vorechovský, I. (2006). Phenotypic consequences of branch point substitutions. Hum. Mutat. *27*, 803–813.

Krawczak, M., Reiss, J., and Cooper, D.N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum. Genet. *90*, 41–54.

Krawczak, M., Thomas, N.S.T., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., and Cooper, D.N. (2007). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum. Mutat. *28*, 150–158.

Kunkel, T.A., and Erie, D.A. (2015). Eukaryotic Mismatch Repair in Relation to DNA Replication. Annu. Rev. Genet. *49*, 291–313.

Kuznetsov, S.G., Liu, P., and Sharan, S.K. (2008). Mouse embryonic stem cell-based functional assay to evaluate mutations in BRCA2. Nat. Med. *14*, 875–881.

Kuznetsov, S.G., Chang, S., and Sharan, S.K. (2010). Functional analysis of human BRCA2 variants using a mouse embryonic stem cell-based assay. Methods Mol. Biol. Clifton NJ *653*, 259–280.

Kwiatkowski, T.J., Bosco, D.A., Leclerc, A.L., Tamrazian, E., Vanderburg, C.R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E.J., Munsat, T., *et al.* (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science *323*, 1205–1208.

Kwok, C.-T., Ward, R.L., Hawkins, N.J., and Hitchins, M.P. (2010). Detection of allelic imbalance in MLH1 expression by pyrosequencing serves as a tool for the identification of germline defects in Lynch syndrome. Fam. Cancer *9*, 345–356.

Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. Int. J. Cell Biol. *2013*, 463786.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Landrum, M.J., and Kattman, B.L. (2018). ClinVar at five years: Delivering on the promise. Hum. Mutat. *39*, 1623–1630.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446*, 926–929.

Lastella, P., Surdo, N.C., Resta, N., Guanti, G., and Stella, A. (2006). *In silico* and in vivo splicing analysis of MLH1 and MSH2 missense mutations shows exon- and tissue-specific effects. BMC Genomics *7*, 243.

Latouche, J.B., and Sadelain, M. (2000). Induction of human cytotoxic T lymphocytes by artificial antigen-presenting cells. Nat. Biotechnol. *18*, 405–409.

Le Hir, H., Izaurralde, E., Maquat, L.E., and Moore, M.J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. *19*, 6860–6869.

Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu. Rev. Biochem. *84*, 291–323.

Lejeune, F., and Maquat, L.E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr. Opin. Cell Biol. *17*, 309–315.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Leman, R., Gaildrat, P., Gac, G.L., Ka, C., Fichou, Y., Audrezet, M.-P., Caux-Moncoutier, V., Caputo, S.M., Boutry-Kryza, N., Léone, M., *et al.* (2018). 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.

Lewandowska, M.A. (2013). The missing puzzle piece: splicing mutations. Int. J. Clin. Exp. Pathol. *6*, 2675–2682.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. U. S. A. *100*, 189–192.

Li, S., and Wilkinson, M.F. (1998). Nonsense surveillance in lymphocytes? Immunity *8*, 135–141.

Li, L., Biswas, K., Habib, L.A., Kuznetsov, S.G., Hamel, N., Kirchhoff, T., Wong, N., Armel, S., Chong, G., Narod, S.A., *et al.* (2009). Functional redundancy of exon 12 of BRCA2 revealed by a comprehensive analysis of the c.6853A>G (p.I2285V) variant. Hum. Mutat. *30*, 1543–1550.

Lin, J.-C. (2017). Therapeutic Applications of Targeted Alternative Splicing to Cancer Treatment. Int. J. Mol. Sci. *19*.

Lin, K.M., Shashidharan, M., Ternent, C.A., Thorson, A.G., Blatchford, G.J., Christensen, M.A., Lanspa, S.J., Lemon, S.J., Watson, P., and Lynch, H.T. (1998). Colorectal and extracolonic cancer

variations in MLH1/MSH2 hereditary nonpolyposis colorectal cancer kindreds and the general population. Dis. Colon Rectum *41*, 428–433.

Lindblom, A., Tannergård, P., Werelius, B., and Nordenskjöld, M. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. Nat. Genet. *5*, 279–282.

Lines, M.A., Huang, L., Schwartzentruber, J., Douglas, S.L., Lynch, D.C., Beaulieu, C., Guion-Almeida, M.L., Zechi-Ceide, R.M., Gener, B., Gillessen-Kaesbach, G., *et al.* (2012). Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. Am. J. Hum. Genet. *90*, 369–377.

Ling, J.P., Pletnikova, O., Troncoso, J.C., and Wong, P.C. (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science *349*, 650–655.

Littink, K.W., Pott, J.-W.R., Collin, R.W.J., Kroes, H.Y., Verheij, J.B.G.M., Blokland, E.A.W., de Castro Miró, M., Hoyng, C.B., Klaver, C.C.W., Koenekoop, R.K., *et al.* (2010). A novel nonsense mutation in CEP290 induces exon skipping and leads to a relatively mild retinal phenotype. Invest. Ophthalmol. Vis. Sci. *51*, 3646–3652.

Liu, F., and Gong, C.-X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol. Neurodegener. *3*, 8.

Liu, S., and Cheng, C. (2013). Alternative RNA splicing and cancer. Wiley Interdiscip. Rev. RNA *4*, 547–566.

Llosa, N.J., Cruise, M., Tam, A., Wicks, E.C., Hechenbleikner, E.M., Taube, J.M., Blosser, R.L., Fan, H., Wang, H., Luber, B.S., *et al.* (2015). The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. Cancer Discov. *5*, 43–51.

Long, J.C., and Caceres, J.F. (2009). The SR protein family of splicing factors: master regulators of gene expression. Biochem. J. *417*, 15–27.

Longman, D., Plasterk, R.H.A., Johnstone, I.L., and Cáceres, J.F. (2007). Mechanistic insights and identification of two novel factors in the C. elegans NMD pathway. Genes Dev. *21*, 1075–1085.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. *579*, 1900–1903.

Losson, R., and Lacroute, F. (1979). Interference of nonsense mutations with eukaryotic messenger RNA stability. Proc. Natl. Acad. Sci. U. S. A. *76*, 5134–5137.

Loveday, C., Turnbull, C., Ruark, E., Xicola, R.M.M., Ramsay, E., Hughes, D., Warren-Perry, M., Snape, K., Breast Cancer Susceptibility Collaboration (UK), Eccles, D., *et al.* (2012). Germline RAD51C mutations confer susceptibility to ovarian cancer. Nat. Genet. *44*, 475–476; author reply 476.

Lykke-Andersen, S., and Jensen, T.H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. Nat. Rev. Mol. Cell Biol. *16*, 665–677.

Lynch, H.T., and Krush, A.J. (1971). Carcinoma of the breast and ovary in three families. Surg. Gynecol. Obstet. *133*, 644–648.

Lynch, H.T., Guirgis, H.A., Albert, S., Brennan, M., Lynch, J., Kraft, C., Pocekay, D., Vaughns, C., and Kaplan, A. (1974). Familial association of carcinoma of the breast and ovary. Surg. Gynecol. Obstet. *138*, 717–724.

Lynch, H.T., Harris, R.E., Guirgis, H.A., Lynch, P.M., Maloney, K., Rankin, L., and Lynch, J. (1976). Early age of onset and familial breast cancer. Lancet Lond. Engl. *2*, 626–627.

Lynch, H.T., Snyder, C., and Casey, M.J. (2013). Hereditary ovarian and breast cancer: what have we learned? Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. *24 Suppl 8*, viii83–viii95.

Lynch, H.T., Snyder, C.L., Shaw, T.G., Heinen, C.D., and Hitchins, M.P. (2015). Milestones of Lynch syndrome: 1895-2015. Nat. Rev. Cancer *15*, 181–194.

Maby, P., Galon, J., and Latouche, J.-B. (2016). Frameshift mutations, neoantigens and tumor-specific CD8(+) T cells in microsatellite unstable colorectal cancers. Oncoimmunology *5*, e1115943.

Mangold, E., Pagenstecher, C., Friedl, W., Fischer, H.-P., Merkelbach-Bruse, S., Ohlendorf, M., Friedrichs, N., Aretz, S., Buettner, R., Propping, P., *et al.* (2005). Tumours from MSH2 mutation carriers show loss of MSH2 expression but many tumours from MLH1 mutation carriers exhibit weak positive MLH1 staining. J. Pathol. *207*, 385–395.

Manley, J.L., and Krainer, A.R. (2010). A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev. *24*, 1073–1074.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature *470*, 198–203.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Marquis, J., Meyer, K., Angehrn, L., Kämpfer, S.S., Rothen-Rutishauser, B., and Schümperli, D. (2007). Spinal muscular atrophy: SMN2 pre-mRNA splicing corrected by a U7 snRNA derivative carrying a splicing enhancer sequence. Mol. Ther. J. Am. Soc. Gene Ther. *15*, 1479–1486.

Martin, S.A., McCarthy, A., Barber, L.J., Burgess, D.J., Parry, S., Lord, C.J., and Ashworth, A. (2009). Methotrexate induces oxidative DNA damage and is selectively lethal to tumour cells with defects in the DNA mismatch repair gene MSH2. EMBO Mol. Med. *1*, 323–337.

Martin, S.A., McCabe, N., Mullarkey, M., Cummins, R., Burgess, D.J., Nakabeppu, Y., Oka, S., Kay, E., Lord, C.J., and Ashworth, A. (2010). DNA polymerases as potential therapeutic targets

for cancers deficient in the DNA mismatch repair proteins MSH2 or MLH1. Cancer Cell *17*, 235–248.

Martinez, J.S., Baldeyron, C., and Carreira, A. (2015). Molding BRCA2 function through its interacting partners. Cell Cycle Georget. Tex *14*, 3389–3395.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.-F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv. Exp. Med. Biol. *623*, 123–147.

Maslen, C., Babcock, D., Raghunath, M., and Steinmann, B. (1997). A rare branch-point mutation is associated with missplicing of fibrillin-2 in a large family with congenital contractural arachnodactyly. Am. J. Hum. Genet. *60*, 1389–1398.

Matsuzawa, A., Kanno, S.-I., Nakayama, M., Mochiduki, H., Wei, L., Shimaoka, T., Furukawa, Y., Kato, K., Shibata, S., Yasui, A., *et al.* (2014). The BRCA1/BARD1-interacting protein OLA1 functions in centrosome regulation. Mol. Cell *53*, 101–114.

Mavaddat, N., Peock, S., Frost, D., Ellis, S., Platte, R., Fineberg, E., Evans, D.G., Izatt, L., Eeles, R.A., Adlard, J., *et al.* (2013). Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. J. Natl. Cancer Inst. *105*, 812–822.

Mazoyer, S., Dunning, A.M., Serova, O., Dearden, J., Puget, N., Healey, C.S., Gayther, S.A., Mangion, J., Stratton, M.R., Lynch, H.T., *et al.* (1996). A polymorphic stop codon in BRCA2. Nat. Genet. *14*, 253–254.

McCullough, A.J., and Berget, S.M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol. Cell. Biol. *17*, 4562–4571.

McCullough, A.J., and Berget, S.M. (2000). An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. Mol. Cell. Biol. *20*, 9225–9235.

McIlwain, D.R., Pan, Q., Reilly, P.T., Elia, A.J., McCracken, S., Wakeham, A.C., Itie-Youten, A., Blencowe, B.J., and Mak, T.W. (2010). Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. Proc. Natl. Acad. Sci. U. S. A. *107*, 12186–12191.

McLornan, D.P., List, A., and Mufti, G.J. (2014). Applying synthetic lethality for the selective targeting of cancer. N. Engl. J. Med. *371*, 1725–1735.

Medghalchi, S.M., Frischmeyer, P.A., Mendell, J.T., Kelly, A.G., Lawler, A.M., and Dietz, H.C. (2001). Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. Hum. Mol. Genet. *10*, 99–105.

Meeks, H.D., Song, H., Michailidou, K., Bolla, M.K., Dennis, J., Wang, Q., Barrowdale, D., Frost, D., EMBRACE, McGuffog, L., *et al.* (2016). BRCA2 Polymorphic Stop Codon K3326X and the Risk of Breast, Prostate, and Ovarian Cancers. J. Natl. Cancer Inst. *108*.

Meindl, A., Hellebrand, H., Wiek, C., Erven, V., Wappenschmidt, B., Niederacher, D., Freund, M., Lichtner, P., Hartmann, L., Schaal, H., *et al.* (2010). Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. Nat. Genet. *42*, 410–414.

Mendell, J.T., Sharifi, N.A., Meyers, J.L., Martinez-Murillo, F., and Dietz, H.C. (2004). Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat. Genet. *36*, 1073–1078.

Menzel, T., Nähse-Kumpf, V., Kousholt, A.N., Klein, D.K., Lund-Andersen, C., Lees, M., Johansen, J.V., Syljuåsen, R.G., and Sørensen, C.S. (2011). A genetic screen identifies BRCA2 and PALB2 as key regulators of G2 checkpoint maintenance. EMBO Rep. *12*, 705–712.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res. *25*, 290–303.

Merkin, J.J., Chen, P., Alexis, M.S., Hautaniemi, S.K., and Burge, C.B. (2015). Origins and impacts of new mammalian exons. Cell Rep. *10*, 1992–2005.

Mesman, R.L.S., Calléja, F.M.G.R., Hendriks, G., Morolli, B., Misovic, B., Devilee, P., van Asperen, C.J., Vrieling, H., and Vreeswijk, M.P.G. (2018). The functional impact of variants of uncertain significance in BRCA2. Genet. Med. Off. J. Am. Coll. Med. Genet.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., and Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66–71.

Milner, J., Ponder, B., Hughes-Davies, L., Seltmann, M., and Kouzarides, T. (1997). Transcriptional activation functions in BRCA2. Nature *386*, 772–773.

Min, J., Choi, E.S., Hwang, K., Kim, J., Sampath, S., Venkitaraman, A.R., and Lee, H. (2012). The breast cancer susceptibility gene BRCA2 is required for the maintenance of telomere homeostasis. J. Biol. Chem. *287*, 5091–5101.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.

Miro, J., Laaref, A.M., Rofidal, V., Lagrafeuille, R., Hem, S., Thorel, D., Méchin, D., Mamchaoui, K., Mouly, V., Claustres, M., *et al.* (2015). FUBP1: a new protagonist in splicing regulation of the DMD gene. Nucleic Acids Res. *43*, 2378–2389.

Mitchell, P., and Tollervey, D. (2000). mRNA stability in eukaryotes. Curr. Opin. Genet. Dev. *10*, 193–198.

Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., Igari, T., Koike, M., Chiba, M., and Mori, T. (1997). Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. Nat. Genet. *17*, 271–272.

Moghadasi, S., Eccles, D.M., Devilee, P., Vreeswijk, M.P.G., and van Asperen, C.J. (2016). Classification and Clinical Management of Variants of Uncertain Significance in High Penetrance Cancer Predisposition Genes. Hum. Mutat. *37*, 331–336.

Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., Santamariña, M., Behar, R., Blanco, A., Carrasco, E., *et al.* (2018). Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? Front. Genet. *9*, 366.

Monteiro, A.N., August, A., and Hanafusa, H. (1996). Evidence for a transcriptional activation function of BRCA1 C-terminal region. Proc. Natl. Acad. Sci. U. S. A. *93*, 13595–13599.

Montejo, J.M., Magallón, M., Tizzano, E., and Solera, J. (1999). Identification of twenty-one new mutations in the factor IX gene by SSCP analysis. Hum. Mutat. *13*, 160–165.

Mordes, D., Luo, X., Kar, A., Kuo, D., Xu, L., Fushimi, K., Yu, G., Sternberg, P., and Wu, J.Y. (2006). Pre-mRNA splicing and retinitis pigmentosa. Mol. Vis. *12*, 1259–1271.

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

Mueller, W.F., Larsen, L.S.Z., Garibaldi, A., Hatfield, G.W., and Hertel, K.J. (2015). The Silent Sway of Splicing by Synonymous Substitutions. J. Biol. Chem. *290*, 27700–27711.

Muir, E.G., Bell, A.J., and Barlow, K.A. (1967). Multiple primary carcinomata of the colon, duodenum, and larynx associated with kerato-acanthomata of the face. Br. J. Surg. *54*, 191–195.

Muller, D., Rouleau, E., Schultz, I., Caputo, S., Lefol, C., Bièche, I., Caron, O., Noguès, C., Limacher, J.M., Demange, L., *et al.* (2011a). An entire exon 3 germ-line rearrangement in the BRCA2 gene: pathogenic relevance of exon 3 deletion in breast cancer predisposition. BMC Med. Genet. *12*, 121.

Muller, D., Rouleau, E., Schultz, I., Caputo, S., Lefol, C., Bièche, I., Caron, O., Noguès, C., Limacher, J.M., Demange, L., *et al.* (2011b). An entire exon 3 germ-line rearrangement in the BRCA2 gene: pathogenic relevance of exon 3 deletion in breast cancer predisposition. BMC Med. Genet. *12*, 121.

Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem. Sci. *23*, 198–199.

Nagy, R., Sweet, K., and Eng, C. (2004). Highly penetrant hereditary cancer syndromes. Oncogene *23*, 6445–6470.

Nakagawa, H., Lockman, J.C., Frankel, W.L., Hampel, H., Steenblock, K., Burgart, L.J., Thibodeau, S.N., and de la Chapelle, A. (2004). Mismatch repair gene PMS2: disease-causing germline mutations are frequent in patients whose tumors stain negative for PMS2 protein, but paralogous genes obscure mutation detection and interpretation. Cancer Res. *64*, 4721–4727.

Naruse, H., Ikawa, N., Yamaguchi, K., Nakamura, Y., Arai, M., Ishioka, C., Sugano, K., Tamura, K., Tomita, N., Matsubara, N., *et al.* (2009). Determination of splice-site mutations in Lynch syndrome (hereditary non-polyposis colorectal cancer) patients using functional splicing assay. Fam. Cancer *8*, 509–517.

Neumann, B., Walter, T., Hériché, J.-K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., *et al.* (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Nicolaides, N.C., Papadopoulos, N., Liu, B., Wei, Y.F., Carter, K.C., Ruben, S.M., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D., and Fraser, C.M. (1994). Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. Nature *371*, 75–80.

Nordling, M., Karlsson, P., Wahlström, J., Engwall, Y., Wallgren, A., and Martinsson, T. (1998). A large deletion disrupts the exon 3 transcription activation domain of the BRCA2 gene in a breast/ovarian cancer family. Cancer Res. *58*, 1372–1375.

Nyström-Lahti, M., Holmberg, M., Fidalgo, P., Salovaara, R., de la Chapelle, A., Jiricny, J., and Peltomäki, P. (1999). Missense and nonsense mutations in codon 659 of MLH1 cause aberrant splicing of messenger RNA in HNPCC kindreds. Genes. Chromosomes Cancer *26*, 372–375.

O'Brien, V., and Brown, R. (2006). Signalling cell cycle arrest and cell death through the MMR System. Carcinogenesis *27*, 682–692.

Ohnishi, T., Yamashita, A., Kashima, I., Schell, T., Anders, K.R., Grimson, A., Hachiya, T., Hentze, M.W., Anderson, P., and Ohno, S. (2003). Phosphorylation of hUPF1 induces formation of mRNA surveillance complexes containing hSMG-5 and hSMG-7. Mol. Cell *12*, 1187–1200.

Olschwang, and Eisinger (2010). Analyse des gènes du MisMatch Repair dans le syndrome de Lynch. Cancéro Dig.

Oltean, S., and Bates, D.O. (2014). Hallmarks of alternative splicing in cancer. Oncogene *33*, 5311–5318.

Omenn, G.S., Menon, R., and Zhang, Y. (2013). Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. J. Proteomics *90*, 28–37.

Ouchi, T., Monteiro, A.N., August, A., Aaronson, S.A., and Hanafusa, H. (1998). BRCA1 regulates p53-dependent gene expression. Proc. Natl. Acad. Sci. U. S. A. *95*, 2302–2306.

Padgett, R.A. (2005). mRNA Splicing: Role of snRNAs. In Encyclopedia of Life Sciences, John Wiley & Sons, Ltd, ed. (Chichester, UK: John Wiley & Sons, Ltd), p.

Padgett, R.A. (2012). New connections between splicing and human disease. Trends Genet. TIG *28*, 147–154.

Padgett, R.A. (2015). Finding the unexpected--how we identified a second class of introns and the U12-dependent spliceosome. RNA N. Y. N *21*, 544–545.

Pagani, F., Buratti, E., Stuani, C., and Baralle, F.E. (2003a). Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. J. Biol. Chem. *278*, 26580–26588.

Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., and Baralle, F.E. (2003b). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. Hum. Mol. Genet. *12*, 1111–1120.

Pagani, F., Raponi, M., and Baralle, F.E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. U. S. A. *102*, 6368–6372.

Pagenstecher, C., Wehner, M., Friedl, W., Rahner, N., Aretz, S., Friedrichs, N., Sengteller, M., Henn, W., Buettner, R., Propping, P., *et al.* (2006). Aberrant splicing in MLH1 and MSH2 due to exonic and intronic variants. Hum. Genet. *119*, 9–22.

Palmirotta, R., Verì, M.C., Curia, M.C., Aceto, G., D'Amico, F., Esposito, D.L., Arcuri, P., Mariani-Costantini, R., Messerini, L., Mori, S., *et al.* (1998). Transcripts with splicings of exons 15 and 16 of the hMLH1 gene in normal lymphocytes: implications in RNA-based mutation screening of hereditary non-polyposis colorectal cancer. Eur. J. Cancer Oxf. Engl. 1990 *34*, 927–930.

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

Papadopoulos, N., Nicolaides, N.C., Wei, Y.F., Ruben, S.M., Carter, K.C., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D., Fraser, C.M., and Adams, M.D. (1994). Mutation of a mutL homolog in hereditary colon cancer. Science *263*, 1625–1629.

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J.S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., *et al.* (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N. Engl. J. Med. *365*, 1384–1395.

Papasaikas, P., and Valcárcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends Biochem. Sci. *41*, 33–45.

Park, D.J., Lesueur, F., Nguyen-Dumont, T., Pertesi, M., Odefrey, F., Hammet, F., Neuhausen, S.L., John, E.M., Andrulis, I.L., Terry, M.B., *et al.* (2012). Rare mutations in XRCC2 increase the risk of breast cancer. Am. J. Hum. Genet. *90*, 734–739.

Parker, R., and Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. Nat. Struct. Mol. Biol. *11*, 121–127.

Patel, A.A., and Steitz, J.A. (2003a). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

Patel, A.A., and Steitz, J.A. (2003b). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742.

Peasland, A., Matheson, E., Hall, A., and Irving, J. (2010). Alternative splicing of hMLH1 in childhood acute lymphoblastic leukaemia and characterisation of the variably expressed Delta9/10 isoform as a dominant negative species. Leuk. Res. *34*, 322–327.

Peltomäki, P. (2001). Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Hum. Mol. Genet. *10*, 735–740.

Peltomäki, P. (2014). Epigenetic mechanisms in the pathogenesis of Lynch syndrome. Clin. Genet. *85*, 403–412.

Peltomäki, P., Aaltonen, L.A., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Green, J.S., Jass, J.R., Weber, J.L., and Leach, F.S. (1993). Genetic mapping of a locus predisposing to human colorectal cancer. Science *260*, 810–812.

Peltz, S.W., Brown, A.H., and Jacobson, A. (1993). mRNA destabilization triggered by premature translational termination depends on at least three cis-acting sequence elements and one trans-acting factor. Genes Dev. *7*, 1737–1754.

Pharoah, P.D., Guilford, P., Caldas, C., and International Gastric Cancer Linkage Consortium (2001). Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology *121*, 1348–1353.

Phuah, S.-Y., Looi, L.-M., Hassan, N., Rhodes, A., Dean, S., Taib, N.A.M., Yip, C.-H., and Teo, S.-H. (2012). Triple-negative breast cancer and PTEN (phosphatase and tensin homologue) loss are predictors of BRCA1 germline mutations in women with early-onset and familial breast cancer, but not in women with isolated late-onset breast cancer. Breast Cancer Res. BCR *14*, R142.

Pilotto, S., Gkountakos, A., Carbognin, L., Scarpa, A., Tortora, G., and Bria, E. (2017). MET exon 14 juxtamembrane splicing mutations: clinical and therapeutical perspectives for cancer therapy. Ann. Transl. Med. *5*, 2.

Piva, F., Giulietti, M., Nocchi, L., and Principato, G. (2009). SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. Bioinforma. Oxf. Engl. *25*, 1211–1213.

Piva, F., Giulietti, M., Burini, A.B., and Principato, G. (2012). SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. Hum. Mutat. *33*, 81–85.

Plaschke, J., Bulitta, C., Saeger, H.D., and Schackert, H.K. (1999). Quantitative differences between aberrant transcripts which occur as common isoforms and due to mutation-based exon skipping of the mismatch repair gene hMLH1. Clin. Chem. Lab. Med. *37*, 883–887.

Plazzer, J.P., Sijmons, R.H., Woods, M.O., Peltomäki, P., Thompson, B., Den Dunnen, J.T., and Macrae, F. (2013). The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. Fam. Cancer *12*, 175–180.

Plon, S.E., Eccles, D.M., Easton, D., Foulkes, W.D., Genuardi, M., Greenblatt, M.S., Hogervorst, F.B.L., Hoogerbrugge, N., Spurdle, A.B., Tavtigian, S.V., *et al.* (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum. Mutat. *29*, 1282–1291.

Pozzoli, U., and Sironi, M. (2005). Silencers regulate both constitutive and alternative splicing events in mammals. Cell. Mol. Life Sci. CMLS *62*, 1579–1604.

Prakash, R., Zhang, Y., Feng, W., and Jasin, M. (2015). Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. Cold Spring Harb. Perspect. Biol. *7*, a016600.

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

Raevaara, T.E., Korhonen, M.K., Lohi, H., Hampel, H., Lynch, E., Lönnqvist, K.E., Holinski-Feder, E., Sutter, C., McKinnon, W., Duraisamy, S., *et al.* (2005). Functional significance and clinical phenotype of nontruncating mismatch repair variants of MLH1. Gastroenterology *129*, 537–549.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature *505*, 302–308.

Rajagopalan, S., Andreeva, A., Rutherford, T.J., and Fersht, A.R. (2010). Mapping the physical and functional interactions between the tumor suppressors p53 and BRCA2. Proc. Natl. Acad. Sci. U. S. A. *107*, 8587–8592.

Ram, O., and Ast, G. (2007). SR proteins: a foot on the exon before the transition from intron to exon definition. Trends Genet. TIG *23*, 5–7.

Ramalho, A.S., Clarke, L.A., Sousa, M., Felicio, V., Barreto, C., Lopes, C., and Amaral, M.D. (2016). Comparative ex vivo, in vitro and *in silico* analyses of a CFTR splicing mutation: Importance of functional studies to establish disease liability of mutations. J. Cyst. Fibros. Off. J. Eur. Cyst. Fibros. Soc. *15*, 21–33.

Raponi, M., Baralle, F.E., and Pagani, F. (2007). Reduced splicing efficiency induced by synonymous substitutions may generate a substrate for natural selection of new splicing isoforms: the case of CFTR exon 12. Nucleic Acids Res. *35*, 606–613.

Raponi, M., Kralovicova, J., Copson, E., Divina, P., Eccles, D., Johnson, P., Baralle, D., and Vorechovsky, I. (2011). Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6. Hum. Mutat. *32*, 436–444.

Raponi, M., Douglas, A.G.L., Tammaro, C., Wilson, D.I., and Baralle, D. (2012). Evolutionary constraint helps unmask a splicing regulatory region in BRCA1 exon 11. PloS One *7*, e37255.

Raponi, M., Smith, L.D., Silipo, M., Stuani, C., Buratti, E., and Baralle, D. (2014). BRCA1 exon 11 a model of long exon splicing regulation. RNA Biol. *11*, 351–359.

Reber, S., Stettler, J., Filosa, G., Colombo, M., Jutzi, D., Lenzken, S.C., Schweingruber, C., Bruggmann, R., Bachi, A., Barabino, S.M., *et al.* (2016). Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. EMBO J. *35*, 1504–1521.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J. Comput. Biol. J. Comput. Mol. Cell Biol. *4*, 311–323.

Reid, S., Schindler, D., Hanenberg, H., Barker, K., Hanks, S., Kalb, R., Neveling, K., Kelly, P., Seal, S., Freund, M., *et al.* (2007). Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nat. Genet. *39*, 162–164.

Rennstam, K., Ringberg, A., Cunliffe, H.E., Olsson, H., Landberg, G., and Hedenfalk, I. (2010). Genomic alterations in histopathologically normal breast tissue from BRCA1 mutation carriers may be caused by BRCA1 haploinsufficiency. Genes. Chromosomes Cancer *49*, 78–90.

Renton, A.E., Chiò, A., and Traynor, B.J. (2014). State of play in amyotrophic lateral sclerosis genetics. Nat. Neurosci. *17*, 17–23.

Reyes, G.X., Schmidt, T.T., Kolodner, R.D., and Hombauer, H. (2015). New insights into the mechanism of DNA mismatch repair. Chromosoma *124*, 443–462.

Rhine, C.L., Cygan, K.J., Soemedi, R., Maguire, S., Murray, M.F., Monaghan, S.F., and Fairbrother, W.G. (2018). Hereditary cancer genes are highly susceptible to splicing mutations. PLoS Genet. *14*, e1007231.

Ribic, C.M., Sargent, D.J., Moore, M.J., Thibodeau, S.N., French, A.J., Goldberg, R.M., Hamilton, S.R., Laurent-Puig, P., Gryfe, R., Shepherd, L.E., *et al.* (2003). Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N. Engl. J. Med. *349*, 247–257.

Ricciardone, M.D., Ozçelik, T., Cevher, B., Ozdağ, H., Tuncer, M., Gürgey, A., Uzunalimoğlu, O., Cetinkaya, H., Tanyeli, A., Erken, E., *et al.* (1999). Human MLH1 deficiency predisposes to hematological malignancy and neurofibromatosis type 1. Cancer Res. *59*, 290–293.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., *et al.* (2015). 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. *17*, 405–424.

Richman, S. (2015). Deficient mismatch repair: Read all about it (Review). Int. J. Oncol. *47*, 1189–1202.

Rodriguez-Bigas, M.A., Boland, C.R., Hamilton, S.R., Henson, D.E., Jass, J.R., Khan, P.M., Lynch, H., Perucho, M., Smyrk, T., Sobin, L., *et al.* (1997). A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. J. Natl. Cancer Inst. *89*, 1758–1762.

Romero, A., García-García, F., López-Perolio, I., Ruiz de Garibay, G., García-Sáenz, J.A., Garre, P., Ayllón, P., Benito, E., Dopazo, J., Díaz-Rubio, E., *et al.* (2015). BRCA1 Alternative splicing landscape in breast tissue samples. BMC Cancer *15*, 219.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Rosenthal, E.T., Bowles, K.R., Pruss, D., van Kan, A., Vail, P.J., McElroy, H., and Wenstrup, R.J. (2015). Exceptions to the rule: case studies in the prediction of pathogenicity for genetic variants in hereditary cancer genes. Clin. Genet. *88*, 533–541.

Roy, R., Chun, J., and Powell, S.N. (2011). BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nat. Rev. Cancer *12*, 68–78.

Rubin, B.Y., and Anderson, S.L. (2008). The molecular basis of familial dysautonomia: overview, new discoveries and implications for directed therapies. Neuromolecular Med. *10*, 148–156.

Ryan, N.J. (2014). Ataluren: first global approval. Drugs *74*, 1709–1714.

Salton, M., Kasprzak, W.K., Voss, T., Shapiro, B.A., Poulikakos, P.I., and Misteli, T. (2015). Inhibition of vemurafenib-resistant melanoma by interference with pre-mRNA splicing. Nat. Commun. *6*, 7103.

Sameer, A.S. (2013). Colorectal cancer: molecular mutations and polymorphisms. Front. Oncol. *3*, 114.

Sameer, A.S., Nissar, S., and Fatima, K. (2014). Mismatch repair pathway: molecules, functions, and role in colorectal carcinogenesis. Eur. J. Cancer Prev. Off. J. Eur. Cancer Prev. Organ. ECP *23*, 246–257.

Sancar, A., Lindsey-Boltz, L.A., Unsal-Kaçmaz, K., and Linn, S. (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annu. Rev. Biochem. *73*, 39–85.

Sandell, L.L., and Zakian, V.A. (1993). Loss of a yeast telomere: arrest, recovery, and chromosome loss. Cell *75*, 729–739.

Sangermano, R., Khan, M., Cornelis, S.S., Richelle, V., Albert, S., Garanto, A., Elmelik, D., Qamar, R., Lugtenberg, D., van den Born, L.I., *et al.* (2018). ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease. Genome Res. *28*, 100–110.

Sanz, D.J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardeñosa, E., Lastra, E., Pagani, F., Miner, C., and Velasco, E.A. (2010). A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *16*, 1957–1967.

Sargent, D.J., Marsoni, S., Monges, G., Thibodeau, S.N., Labianca, R., Hamilton, S.R., French, A.J., Kabat, B., Foster, N.R., Torri, V., *et al.* (2010). Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *28*, 3219–3226.

Savage, K.I., and Harkin, D.P. (2015). BRCA1, a "complex" protein involved in the maintenance of genomic stability. FEBS J. *282*, 630–646.

Savisaar, R., and Hurst, L.D. (2017). Estimating the prevalence of functional exonic splice regulatory information. Hum. Genet. *136*, 1059–1078.

Sawyer, S.L., Tian, L., Kähkönen, M., Schwartzentruber, J., Kircher, M., University of Washington Centre for Mendelian Genomics, FORGE Canada Consortium, Majewski, J., Dyment, D.A., Innes, A.M., *et al.* (2015). Biallelic mutations in BRCA1 cause a new Fanconi anemia subtype. Cancer Discov. *5*, 135–142.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671–684.

Schümperli, D., and Pillai, R.S. (2004). The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein. Cell. Mol. Life Sci. CMLS *61*, 2560–2570.

Schwartz, S., Hall, E., and Ast, G. (2009). SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. Nucleic Acids Res. *37*, W189-192.

Scully, R., Anderson, S.F., Chao, D.M., Wei, W., Ye, L., Young, R.A., Livingston, D.M., and Parvin, J.D. (1997). BRCA1 is a component of the RNA polymerase II holoenzyme. Proc. Natl. Acad. Sci. U. S. A. *94*, 5605–5610.

Sedic, M., and Kuperwasser, C. (2016). BRCA1-hapoinsufficiency: Unraveling the molecular and cellular basis for tissue-specific cancer. Cell Cycle Georget. Tex *15*, 621–627.

Sehgal, R., Sheahan, K., O'Connell, P.R., Hanly, A.M., Martin, S.T., and Winter, D.C. (2014). Lynch syndrome: an updated review. Genes *5*, 497–507.

Senapathy, P., Shapiro, M.B., and Harris, N.L. (1990). Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol. *183*, 252–278.

Seo, A., Steinberg-Shemer, O., Unal, S., Casadei, S., Walsh, T., Gumruk, F., Shalev, S., Shimamura, A., Akarsu, N.A., Tamary, H., *et al.* (2018). Mechanism for survival of homozygous nonsense mutations in the tumor suppressor gene *BRCA1*. Proc. Natl. Acad. Sci. *115*, 5241–5246.

Service, R.F. (2006). Gene sequencing. The race for the $1000 genome. Science *311*, 1544–1546.

Sevcik, J., Falk, M., Kleiblova, P., Lhota, F., Stefancikova, L., Janatova, M., Weiterova, L., Lukasova, E., Kozubek, S., Pohlreich, P., *et al.* (2012). The BRCA1 alternative splicing variant Δ14-15 with an in-frame deletion of part of the regulatory serine-containing domain (SCD) impairs the DNA repair capacity in MCF-7 cells. Cell. Signal. *24*, 1023–1030.

Sevcik, J., Falk, M., Macurek, L., Kleiblova, P., Lhota, F., Hojny, J., Stefancikova, L., Janatova, M., Bartek, J., Stribrna, J., *et al.* (2013). Expression of human BRCA1Δ17-19 alternative splicing variant with a truncated BRCT domain in MCF-7 cells results in impaired assembly of DNA repair complexes and aberrant DNA damage response. Cell. Signal. *25*, 1186–1193.

Shao, N., Chai, Y.L., Shyam, E., Reddy, P., and Rao, V.N. (1996). Induction of apoptosis by the tumor suppressor protein BRCA1. Oncogene *13*, 1–7.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155–7174.

Sharan, S.K., Morimatsu, M., Albrecht, U., Lim, D.S., Regel, E., Dinh, C., Sands, A., Eichele, G., Hasty, P., and Bradley, A. (1997). Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. Nature *386*, 804–810.

Sharma, N., Sosnay, P.R., Ramalho, A.S., Douville, C., Franca, A., Gottschalk, L.B., Park, J., Lee, M., Vecchio-Pagan, B., Raraigh, K.S., *et al.* (2014). Experimental assessment of splicing variants using expression minigenes and comparison with *in silico* predictions. Hum. Mutat. *35*, 1249–1259.

Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C., and Black, D.L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat. Struct. Mol. Biol. *15*, 183–191.

Sharp, P.A. (2005). The discovery of split genes and RNA splicing. Trends Biochem. Sci. *30*, 279–281.

Sharp, A., Pichert, G., Lucassen, A., and Eccles, D. (2004). RNA analysis reveals splicing mutations and loss of expression defects in MLH1 and BRCA1. Hum. Mutat. *24*, 272.

Shchepachev, V., Wischnewski, H., Missiaglia, E., Soneson, C., and Azzalin, C.M. (2012). Mpn1, mutated in poikiloderma with neutropenia protein 1, is a conserved 3'-to-5' RNA exonuclease processing U6 small nuclear RNA. Cell Rep. *2*, 855–865.

Shen, M., and Mattox, W. (2012). Activation and repression functions of an SR splicing regulator depend on exonic versus intronic-binding position. Nucleic Acids Res. *40*, 428–437.

Shendure, J. (2011). Next-generation human genetics. Genome Biol. *12*, 408.

Sheng, J.-Q., Zhang, H., Ji, M., Fu, L., Mu, H., Zhang, M.-Z., Huang, J.-S., Han, M., Li, A.-Q., Wei, Z., *et al.* (2009). Genetic diagnosis strategy of hereditary non-polyposis colorectal cancer. World J. Gastroenterol. *15*, 983–989.

Shepard, P.J., and Hertel, K.J. (2009). The SR protein family. Genome Biol. *10*, 242.

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. *34*, 3955–3967.

Shimelis, H., Mesman, R.L.S., Von Nicolai, C., Ehlen, A., Guidugli, L., Martin, C., Calléja, F.M.G.R., Meeks, H., Hallberg, E., Hinton, J., *et al.* (2017). BRCA2 Hypomorphic Missense Variants Confer Moderate Risks of Breast Cancer. Cancer Res. *77*, 2789–2799.

Shoemaker, C.J., and Green, R. (2012). Translation drives mRNA quality control. Nat. Struct. Mol. Biol. *19*, 594–601.

Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. Nat. Rev. Genet. *17*, 407–421.

Siegel, R.L., Miller, K.D., Fedewa, S.A., Ahnen, D.J., Meester, R.G.S., Barzi, A., and Jemal, A. (2017). Colorectal cancer statistics, 2017. CA. Cancer J. Clin. *67*, 177–193.

Signal, B., Gloss, B.S., Dinger, M.E., and Mercer, T.R. (2018). Machine learning annotation of human branchpoints. Bioinforma. Oxf. Engl. *34*, 920–927.

Singh, B., and Eyras, E. (2017). The role of alternative splicing in cancer. Transcription *8*, 91–98.

Singh, R.K., and Cooper, T.A. (2012). Pre-mRNA splicing in disease and therapeutics. Trends Mol. Med. *18*, 472–482.

Singh, N.N., Androphy, E.J., and Singh, R.N. (2004a). An extended inhibitory context causes skipping of exon 7 of SMN2 in spinal muscular atrophy. Biochem. Biophys. Res. Commun. *315*, 381–388.

Singh, N.N., Androphy, E.J., and Singh, R.N. (2004b). In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. RNA N. Y. N *10*, 1291–1305.

Singh, N.N., Singh, R.N., and Androphy, E.J. (2007). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Res. *35*, 371–389.

Slaugenhaupt, S.A., Blumenfeld, A., Gill, S.P., Leyne, M., Mull, J., Cuajungco, M.P., Liebert, C.B., Chadwick, B., Idelson, M., Reznik, L., *et al.* (2001). Tissue-specific expression of a splicing mutation in the IKBKAP gene causes familial dysautonomia. Am. J. Hum. Genet. *68*, 598–605.

Smith, E.C. (2012). An overview of hereditary breast and ovarian cancer syndrome. J. Midwifery Womens Health *57*, 577–584.

Smith, P.J., Zhang, C., Wang, J., Chew, S.L., Zhang, M.Q., and Krainer, A.R. (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum. Mol. Genet. *15*, 2490–2508.

So, B.R., Wan, L., Zhang, Z., Li, P., Babiash, E., Duan, J., Younis, I., and Dreyfuss, G. (2016). A U1 snRNP-specific assembly pathway reveals the SMN complex as a versatile hub for RNP exchange. Nat. Struct. Mol. Biol. *23*, 225–230.

Sonnenblick, A., de Azambuja, E., Azim, H.A., and Piccart, M. (2015). An update on PARP inhibitors--moving to the adjuvant setting. Nat. Rev. Clin. Oncol. *12*, 27–41.

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., Tosi, M., and Martins, A. (2016). Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using *In silico* Tools. PLoS Genet. *12*, e1005756.

Southey, M.C., Ramus, S.J., Dowty, J.G., Smith, L.D., Tesoriero, A.A., Wong, E.E.M., Dite, G.S., Jenkins, M.A., Byrnes, G.B., Winship, I., *et al.* (2011). Morphological predictors of BRCA1 germline mutations in young women with breast cancer. Br. J. Cancer *104*, 903–909.

Spies, M., and Fishel, R. (2015). Mismatch repair during homologous and homeologous recombination. Cold Spring Harb. Perspect. Biol. *7*, a022657.

Spitali, P., and Aartsma-Rus, A. (2012). Splice modulating therapies for human disease. Cell *148*, 1085–1088.

Spurdle, A.B., Couch, F.J., Hogervorst, F.B.L., Radice, P., Sinilnikova, O.M., and IARC Unclassified Genetic Variants Working Group (2008). Prediction and assessment of splicing alterations: implications for clinical testing. Hum. Mutat. *29*, 1304–1313.

Spurdle, A.B., Healey, S., Devereau, A., Hogervorst, F.B.L., Monteiro, A.N.A., Nathanson, K.L., Radice, P., Stoppa-Lyonnet, D., Tavtigian, S., Wappenschmidt, B., *et al.* (2012). 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. *33*, 2–7.

Spurdle, A.B., Couch, F.J., Parsons, M.T., McGuffog, L., Barrowdale, D., Bolla, M.K., Wang, Q., Healey, S., Schmutzler, R., Wappenschmidt, B., *et al.* (2014). Refined histopathological predictors

of BRCA1 and BRCA2 mutation status: a large-scale analysis of breast cancer characteristics from the BCAC, CIMBA, and ENIGMA consortia. Breast Cancer Res. BCR *16*, 3419.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005). Function of alternative splicing. Gene *344*, 1–20.

Starita, L.M., Machida, Y., Sankaran, S., Elias, J.E., Griffin, K., Schlegel, B.P., Gygi, S.P., and Parvin, J.D. (2004). BRCA1-dependent ubiquitination of gamma-tubulin regulates centrosome number. Mol. Cell. Biol. *24*, 8457–8466.

Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. Am. J. Hum. Genet. *101*, 315–325.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T. vO (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. Eur. J. Hum. Genet. EJHG *22*, 1362–1368.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.

Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N., and Sanford, J.R. (2011). Loss of exon identity is a common mechanism of human inherited disease. Genome Res. *21*, 1563–1571.

Sullivan-Reed, K., Bolton-Gillespie, E., Dasgupta, Y., Langer, S., Siciliano, M., Nieborowska-Skorska, M., Hanamshet, K., Belyaeva, E.A., Bernhardy, A.J., Lee, J., *et al.* (2018). Simultaneous Targeting of PARP1 and RAD52 Triggers Dual Synthetic Lethality in BRCA-Deficient Tumor Cells. Cell Rep. *23*, 3127–3136.

Sumpter, R., Sirasanagandla, S., Fernández, Á.F., Wei, Y., Dong, X., Franco, L., Zou, Z., Marchal, C., Lee, M.Y., Clapp, D.W., *et al.* (2016). Fanconi Anemia Proteins Function in Mitophagy and Immunity. Cell *165*, 867–881.

Sun, H., and Chasin, L.A. (2000). Multiple Splicing Defects in an Intronic False Exon. Mol. Cell. Biol. *20*, 6414–6425.

Suñé-Pou, M., Prieto-Sánchez, S., Boyero-Corral, S., Moreno-Castro, C., El Yousfi, Y., Suñé-Negre, J.M., Hernández-Munain, C., and Suñé, C. (2017). Targeting Splicing in the Treatment of Human Disease. Genes *8*.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.

Sureau, A., Gattoni, R., Dooghe, Y., Stévenin, J., and Soret, J. (2001). SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. EMBO J. *20*, 1785–1796.

Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R.A., and Skotheim, R.I. (2016). Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene *35*, 2413–2427.

Taggart, A.J., DeSimone, A.M., Shih, J.S., Filloux, M.E., and Fairbrother, W.G. (2012). Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nat. Struct. Mol. Biol. *19*, 719–721.

Taggart, A.J., Lin, C.-L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. Genome Res. *27*, 639–649.

Tai, Y.C., Domchek, S., Parmigiani, G., and Chen, S. (2007). Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. J. Natl. Cancer Inst. *99*, 1811–1814.

Tajnik, M., Rogalska, M.E., Bussani, E., Barbon, E., Balestra, D., Pinotti, M., and Pagani, F. (2016). Molecular Basis and Therapeutic Strategies to Rescue Factor IX Variants That Affect Splicing and Protein Function. PLoS Genet. *12*, e1006082.

Takahashi, R., and Nagai, K. (2009). Differences in expression between transcripts using alternative promoters of hMLH1 gene and their correlation with microsatellite instability. Oncol. Rep. *22*, 265–271.

Takaoka, M., and Miki, Y. (2018). BRCA1 gene: function and deficiency. Int. J. Clin. Oncol. *23*, 36–44.

Tanko, Q., Franklin, B., Lynch, H., and Knezetic, J. (2002). A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. Mutat. Res. *503*, 37–42.

Teraoka, S.N., Telatar, M., Becker-Catania, S., Liang, T., Onengüt, S., Tolun, A., Chessa, L., Sanal, O., Bernatowska, E., Gatti, R.A., *et al.* (1999). Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am. J. Hum. Genet. *64*, 1617–1631.

Thanaraj, T.A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res. *29*, 2581–2593.

Théry, J.C., Krieger, S., Gaildrat, P., Révillion, F., Buisine, M.-P., Killian, A., Duponchel, C., Rousselin, A., Vaur, D., Peyrat, J.-P., *et al.* (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. Eur. J. Hum. Genet. *19*, 1052–1058.

Thirthagiri, E., Klarmann, K.D., Shukla, A.K., Southon, E., Biswas, K., Martin, B.K., North, S.L., Magidson, V., Burkett, S., Haines, D.C., *et al.* (2016). BRCA2 minor transcript lacking exons 4-7 supports viability in mice and may account for survival of humans with a pathogenic biallelic mutation. Hum. Mol. Genet. *25*, 1934–1945.

Thomassen, M., Blanco, A., Montagna, M., Hansen, T.V.O., Pedersen, I.S., Gutiérrez-Enríquez, S., Menéndez, M., Fachal, L., Santamariña, M., Steffensen, A.Y., *et al.* (2012). Characterization of

BRCA1 and BRCA2 splicing variants: a collaborative report by ENIGMA consortium members. Breast Cancer Res. Treat. *132*, 1009–1023.

Thompson, B.A., Spurdle, A.B., Plazzer, J.-P., Greenblatt, M.S., Akagi, K., Al-Mulla, F., Bapat, B., Bernstein, I., Capellá, G., den Dunnen, J.T., *et al.* (2014). Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat. Genet. *46*, 107–115.

Thompson, B.A., Martins, A., and Spurdle, A.B. (2015). A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays. Clin. Genet. *87*, 100–108.

Tollervey, D. (2006). Molecular biology: RNA lost in translation. Nature *440*, 425–426.

Torre, D. (1968). Multiple sebaceous tumors. Arch. Dermatol. *98*, 549–551.

Tournier, I., Raux, G., Di Fiore, F., Maréchal, I., Leclerc, C., Martin, C., Wang, Q., Buisine, M.-P., Stoppa-Lyonnet, D., Olschwang, S., *et al.* (2004). Analysis of the allele-specific expression of the mismatch repair gene MLH1 using a simple DHPLC-Based Method. Hum. Mutat. *23*, 379–384.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J., *et al.* (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. *29*, 1412–1424.

Trojan, J., Zeuzem, S., Randolph, A., Hemmerle, C., Brieger, A., Raedle, J., Plotz, G., Jiricny, J., and Marra, G. (2002). Functional analysis of hMLH1 variants and HNPCC-related mutations using a human expression system. Gastroenterology *122*, 211–219.

Tuffery-Giraud, S., Saquet, C., Thorel, D., Disset, A., Rivier, F., Malcolm, S., and Claustres, M. (2005). Mutation spectrum leading to an attenuated phenotype in dystrophinopathies. Eur. J. Hum. Genet. EJHG *13*, 1254–1260.

Turcot, J., Despres, J.P., and St Pierre, F. (1959). Malignant tumors of the central nervous system associated with familial polyposis of the colon: report of two cases. Dis. Colon Rectum *2*, 465–468.

Turunen, J.J., Niemelä, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA *4*, 61–76.

Unger, M.A., Nathanson, K.L., Calzone, K., Antin-Ozerkis, D., Shih, H.A., Martin, A.M., Lenoir, G.M., Mazoyer, S., and Weber, B.L. (2000). Screening for genomic rearrangements in families with breast and ovarian cancer identifies BRCA1 mutations previously missed by conformation-sensitive gel electrophoresis or sequencing. Am. J. Hum. Genet. *67*, 841–850.

Vance, C., Rogelj, B., Hortobágyi, T., De Vos, K.J., Nishimura, A.L., Sreedharan, J., Hu, X., Smith, B., Ruddy, D., Wright, P., *et al.* (2009). Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. Science *323*, 1208–1211.

Vasen, H.F.A. (2007). Review article: The Lynch syndrome (hereditary nonpolyposis colorectal cancer). Aliment. Pharmacol. Ther. *26 Suppl 2*, 113–126.

Vasen, H.F.A., and de Vos Tot Nederveen Cappel, W.H. (2013). A hundred years of Lynch syndrome research (1913-2013). Fam. Cancer *12*, 141–142.

Vasen, H.F., Watson, P., Mecklin, J.P., and Lynch, H.T. (1999). New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. Gastroenterology *116*, 1453–1456.

Vasen, H.F., Stormorken, A., Menko, F.H., Nagengast, F.M., Kleibeuker, J.H., Griffioen, G., Taal, B.G., Moller, P., and Wijnen, J.T. (2001). MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *19*, 4074–4080.

Vasudevan, S., Peltz, S.W., and Wilusz, C.J. (2002). Non-stop decay--a new mRNA surveillance pathway. BioEssays News Rev. Mol. Cell. Dev. Biol. *24*, 785–788.

Vaughn, C.P., Robles, J., Swensen, J.J., Miller, C.E., Lyon, E., Mao, R., Bayrak-Toydemir, P., and Samowitz, W.S. (2010). Clinical analysis of PMS2: mutation detection and avoidance of pseudogenes. Hum. Mutat. *31*, 588–593.

Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. Hum. Genet. *136*, 1093–1111.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The sequence of the human genome. Science *291*, 1304–1351.

Vilenchik, M.M., and Knudson, A.G. (2003). Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. Proc. Natl. Acad. Sci. U. S. A. *100*, 12871–12876.

Vogelstein, B., and Kinzler, K.W. (2015). The Path to Cancer --Three Strikes and You’re Out. N. Engl. J. Med. *373*, 1895–1898.

Volard, B., Krieger, S., Planchard, G., Hardouin, A., Vaur, D., Rame, J.-P., and Bardet, S. (2012). Assessment of SPAG9 Transcript in Fine Needle Aspirates of Thyroid Nodules. Eur. Thyroid J. *1*, 118–121.

Vorechovsky, I. (2010). Transposable elements in disease-associated cryptic exons. Hum. Genet. *127*, 135–154.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701–718.

Walker, L.C., Whiley, P.J., Houdayer, C., Hansen, T.V.O., Vega, A., Santamarina, M., Blanco, A., Fachal, L., Southey, M.C., Lafferty, A., *et al.* (2013). 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. *34*, 1424–1431.

Wally, V., Murauer, E.M., and Bauer, J.W. (2012). Spliceosome-mediated trans-splicing: the therapeutic cut and paste. J. Invest. Dermatol. *132*, 1959–1966.

Walsh, T., Casadei, S., Coats, K.H., Swisher, E., Stray, S.M., Higgins, J., Roach, K.C., Mandell, J., Lee, M.K., Ciernikova, S., *et al.* (2006). Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA *295*, 1379–1388.

Wan, L., and Dreyfuss, G. (2017). Splicing-Correcting Therapy for SMA. Cell *170*, 5.

Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. *8*, 749–761.

Wang, Z., and Burge, C.B. (2008a). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA N. Y. N *14*, 802–813.

Wang, Z., and Burge, C.B. (2008b). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA N. Y. N *14*, 802–813.

Wang, M., Price, C., Han, J., Cisler, J., Imaizumi, K., Van Thienen, M.N., DePaepe, A., and Godfrey, M. (1995). Recurrent mis-splicing of fibrillin exon 32 in two patients with neonatal Marfan syndrome. Hum. Mol. Genet. *4*, 607–613.

Wang, Q., Lasset, C., Desseigne, F., Frappaz, D., Bergeron, C., Navarro, C., Ruano, E., and Puisieux, A. (1999). Neurofibromatosis and early onset of cancers in hMLH1-deficient children. Cancer Res. *59*, 294–297.

Wang, X., Xue, C., Wang, X., Liu, H., Xu, Y., Zhao, R., Jiang, Z., Dodson, M.V., and Chen, J. (2009a). Differential display of expressed genes reveals a novel function of SFRS18 in regulation of intramuscular fat deposition. Int. J. Biol. Sci. *5*, 28–33.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831–845.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Wappenschmidt, B., Becker, A.A., Hauke, J., Weber, U., Engert, S., Köhler, J., Kast, K., Arnold, N., Rhiem, K., Hahnen, E., *et al.* (2012). 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 *7*, e50800.

Ward, A.J., and Cooper, T.A. (2010). The pathobiology of splicing. J. Pathol. *220*, 152–163.

Warthin, A.S. (1985). Classics in oncology. Heredity with reference to carcinoma as shown by the study of the cases examined in the pathological laboratory of the University of Michigan, 1895-1913. CA. Cancer J. Clin. *35*, 348–359.

Weischenfeldt, J., Damgaard, I., Bryder, D., Theilgaard-Mönch, K., Thoren, L.A., Nielsen, F.C., Jacobsen, S.E.W., Nerlov, C., and Porse, B.T. (2008). NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev. *22*, 1381–1396.

Welcsh, P.L., and King, M.C. (2001). BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. Hum. Mol. Genet. *10*, 705–713.

Westdorp, H., Fennemann, F.L., Weren, R.D.A., Bisseling, T.M., Ligtenberg, M.J.L., Figdor, C.G., Schreibelt, G., Hoogerbrugge, N., Wimmers, F., and de Vries, I.J.M. (2016). Opportunities for immunotherapy in microsatellite instable colorectal cancer. Cancer Immunol. Immunother. *65*, 1249–1259.

Whiley, P.J., de la Hoya, M., Thomassen, M., Becker, A., Brandão, R., Pedersen, I.S., Montagna, M., Menéndez, M., Quiles, F., Gutiérrez-Enríquez, S., *et al.* (2014). Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. Clin. Chem. *60*, 341–352.

Will, C.L., and Lührmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. Biol. Chem. *386*, 713–724.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. *3*.

de Winter, J.P., and Joenje, H. (2009). The genetic and molecular basis of Fanconi anemia. Mutat. Res. *668*, 11–19.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789–792.

Xia, B., Dorsman, J.C., Ameziane, N., de Vries, Y., Rooimans, M.A., Sheng, Q., Pals, G., Errami, A., Gluckman, E., Llera, J., *et al.* (2007). Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. Nat. Genet. *39*, 159–161.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yamazaki, T., Chen, S., Yu, Y., Yan, B., Haertlein, T.C., Carrasco, M.A., Tapia, J.C., Zhai, B., Das, R., Lalancette-Hebert, M., *et al.* (2012). FUS-SMN protein interactions link the motor neuron diseases ALS and SMA. Cell Rep. *2*, 799–806.

Yan, H., Papadopoulos, N., Marra, G., Perrera, C., Jiricny, J., Boland, C.R., Lynch, H.T., Chadwick, R.B., de la Chapelle, A., Berg, K., *et al.* (2000). Conversion of diploidy to haploidy. Nature *403*, 723–724.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. J. Comput. Mol. Cell Biol. *11*, 377–394.

Yeo, G., Holste, D., Kreiman, G., and Burge, C.B. (2004). Variation in alternative splicing across human tissues. Genome Biol. *5*, R74.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., *et al.* (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69.

Yu, X., and Chen, J. (2004). DNA damage-induced cell cycle checkpoint control requires CtIP, a phosphorylation-dependent binding partner of BRCA1 C-terminal domains. Mol. Cell. Biol. *24*, 9478–9486.

Zainal Abidin, N., Haq, I.J., Gardner, A.I., and Brodlie, M. (2017). Ataluren in cystic fibrosis: development, clinical studies and where are we now? Expert Opin. Pharmacother. *18*, 1363–1371.

Zhang, X.H.-F., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. *18*, 1241–1250.

Zhang, C., Li, W.-H., Krainer, A.R., and Zhang, M.Q. (2008). RNA landscape of evolution for optimal exon and intron discrimination. Proc. Natl. Acad. Sci. U. S. A. *105*, 5797–5802.

Zhang, J., Lieu, Y.K., Ali, A.M., Penson, A., Reggio, K.S., Rabadan, R., Raza, A., Mukherjee, S., and Manley, J.L. (2015). Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. Proc. Natl. Acad. Sci. U. S. A. *112*, E4726-4734.

Zhang, Q., Fan, X., Wang, Y., Sun, M.-A., Shao, J., and Guo, D. (2017). BPP: a sequence-based algorithm for branch point prediction. Bioinforma. Oxf. Engl. *33*, 3166–3172.

Zhang, X., Moréra, S., Bates, P.A., Whitehead, P.C., Coffer, A.I., Hainbucher, K., Nash, R.A., Sternberg, M.J., Lindahl, T., and Freemont, P.S. (1998). Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. EMBO J. *17*, 6404–6411.

Zhu, Q., Pao, G.M., Huynh, A.M., Suh, H., Tonnu, N., Nederlof, P.M., Gage, F.H., and Verma, I.M. (2011). BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. Nature *477*, 179–184.

Zou, J.P., Hirose, Y., Siddique, H., Rao, V.N., and Reddy, E.S. (1999). Structure and expression of variant BRCA2a lacking the transactivation domain. Oncol. Rep. *6*, 437–440.