Université de Lille
Année Universitaire 2017/2018

Faculté des Sciences Pharmaceutiques et Biologiques de Lille

MEMOIRE
POUR LE DIPLOME D'ETUDES SPECIALISEES
DE BIOLOGIE MEDICALE

Soutenu publiquement le 16 Février 2018
Par M. GRZYCH Guillaume

Conformément aux dispositions du Décret du 10 septembre 1990
tient lieu de
THESE EN VUE DU DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Evaluation des outils de prédiction *in silico* et intérêt des tests fonctionnels dans l'interprétation des variants identifiés par séquençage de nouvelle génération en génétique humaine.

Membres du jury :

Président :	Monsieur le Professeur Thierry BROUSSEAU Faculté de Pharmacie, Université de Lille II
Assesseur :	Madame le Professeur Marie-Pierre BUISINE Faculté de Médecine, Université de Lille II
Directeurs de mémoire :	Madame le Docteur Julie LECLERC Faculté de Médecine, Université de Lille II
	Monsieur le Docteur Jamal GHOUMID Faculté de Médecine, Université de Lille II

Faculté des Sciences Pharmaceutiques
et Biologiques de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
Tel. : 03.20.96.40.40 - Télécopie : 03.20.96.43.64
http://pharmacie.univ-lille2.fr

Faculté de Pharmacie
de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
☎ 03.20.96.40.40 - 🖷 : 03.20.96.43.64

Université de Lille

Président :	Jean-Christophe CAMART
Premier Vice-président :	Damien CUNY
Vice-présidente Formation :	Lynne FRANJIÉ
Vice-président Recherche :	Lionel MONTAGNE
Vice-président Relations Internationales :	François-Olivier SEYS
Directeur Général des Services :	Pierre-Marie ROBERT
Directrice Générale des Services Adjointe :	Marie-Dominique SAVINA

Faculté de Pharmacie

Doyen :	Bertrand DÉCAUDIN
Vice-Doyen et Assesseur à la Recherche :	Patricia MELNYK
Assesseur aux Relations Internationales :	Philippe CHAVATTE
Assesseur à la Vie de la Faculté et aux Relations avec le Monde Professionnel :	Thomas MORGENROTH
Assesseur à la Pédagogie :	Benjamin BERTIN
Assesseur à la Scolarité :	Christophe BOCHU
Responsable des Services :	Cyrille PORTA

Liste des Professeurs des Universités - Praticiens Hospitaliers

Civ.	NOM	Prénom	Laboratoire
Mme	ALLORGE	Delphine	Toxicologie
M.	BROUSSEAU	Thierry	Biochimie
M.	DÉCAUDIN	Bertrand	Pharmacie Galénique
M.	DEPREUX	Patrick	ICPAL
M.	DINE	Thierry	Pharmacie clinique
Mme	DUPONT-PRADO	Annabelle	Hématologie
M.	GRESSIER	Bernard	Pharmacologie
M.	LUYCKX	Michel	Pharmacie clinique
M.	ODOU	Pascal	Pharmacie Galénique
M.	STAELS	Bart	Biologie Cellulaire

Liste des Professeurs des Universités

Civ.	NOM	Prénom	Laboratoire
M.	ALIOUAT	El Moukhtar	Parasitologie
Mme	AZAROUAL	Nathalie	Physique
M.	BERTHELOT	Pascal	Onco et Neurochimie
M.	CAZIN	Jean-Louis	Pharmacologie – Pharmacie clinique
M.	CHAVATTE	Philippe	ICPAL
M.	COURTECUISSE	Régis	Sciences végétales et fongiques
M.	CUNY	Damien	Sciences végétales et fongiques
Mme	DELBAERE	Stéphanie	Physique
M.	DEPREZ	Benoît	Lab. de Médicaments et Molécules
Mme	DEPREZ	Rebecca	Lab. de Médicaments et Molécules
M.	DUPONT	Frédéric	Sciences végétales et fongiques
M.	DURIEZ	Patrick	Physiologie
M.	FOLIGNE	Benoît	Bactériologie
M.	GARÇON	Guillaume	Toxicologie
Mme	GAYOT	Anne	Pharmacotechnie Industrielle

M.	GOOSSENS	Jean François	Chimie Analytique
M.	HENNEBELLE	Thierry	Pharmacognosie
M.	LEMDANI	Mohamed	Biomathématiques
Mme	LESTAVEL	Sophie	Biologie Cellulaire
M.	LUC	Gerald	Physiologie
Mme	MELNYK	Patricia	Onco et Neurochimie
M.	MILLET	Régis	ICPAL
Mme	MUHR – TAILLEUX	Anne	Biochimie
Mme	PAUMELLE-LESTRELIN	Réjane	Biologie Cellulaire
Mme	PERROY	Anne Catherine	Législation
Mme	ROMOND	Marie Bénédicte	Bactériologie
Mme	SAHPAZ	Sevser	Pharmacognosie
M.	SERGHERAERT	Eric	Législation
Mme	SIEPMANN	Florence	Pharmacotechnie Industrielle
M.	SIEPMANN	Juergen	Pharmacotechnie Industrielle
M.	WILLAND	Nicolas	Lab. de Médicaments et Molécules

Liste des Maîtres de Conférences - Praticiens Hospitaliers

Civ.	NOM	Prénom	Laboratoire
Mme	BALDUYCK	Malika	Biochimie
Mme	GARAT	Anne	Toxicologie
Mme	GOFFARD	Anne	Bactériologie
M.	LANNOY	Damien	Pharmacie Galénique
Mme	ODOU	Marie Françoise	Bactériologie
M.	SIMON	Nicolas	Pharmacie Galénique

Liste des Maîtres de Conférences

Civ.	NOM	Prénom	Laboratoire
Mme	ALIOUAT	Cécile Marie	Parasitologie
M.	ANTHERIEU	Sébastien	Toxicologie

Mme	AUMERCIER	Pierrette	Biochimie
Mme	BANTUBUNGI	Kadiombo	Biologie cellulaire
Mme	BARTHELEMY	Christine	Pharmacie Galénique
Mme	BEHRA	Josette	Bactériologie
M	BELARBI	Karim	Pharmacologie
M.	BERTHET	Jérôme	Physique
M.	BERTIN	Benjamin	Immunologie
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle
M.	BOCHU	Christophe	Physique
M.	BORDAGE	Simon	Pharmacognosie
M.	BOSC	Damien	Lab. de Médicaments et Molécules
M.	BRIAND	Olivier	Biochimie
M.	CARNOY	Christophe	Immunologie
Mme	CARON	Sandrine	Biologie cellulaire
Mme	CHABÉ	Magali	Parasitologie
Mme	CHARTON	Julie	Lab. de Médicaments et Molécules
M	CHEVALIER	Dany	Toxicologie
M.	COCHELARD	Dominique	Biomathématiques
Mme	DANEL	Cécile	Chimie Analytique
Mme	DEMANCHE	Christine	Parasitologie
Mme	DEMARQUILLY	Catherine	Biomathématiques
M.	DHIFLI	Wajdi	Biomathématiques
Mme	DUMONT	Julie	Biologie cellulaire
Mme	DUTOUT-AGOURIDAS	Laurence	Onco et Neurochimie
M.	EL BAKALI	Jamal	Onco et Neurochimie
M.	FARCE	Amaury	ICPAL
Mme	FLIPO	Marion	Lab. de Médicaments et Molécules
Mme	FOULON	Catherine	Chimie Analytique
M.	FURMAN	Christophe	ICPAL
Mme	GENAY	Stéphanie	Pharmacie Galénique
M.	GERVOIS	Philippe	Biochimie

Mme	GOOSSENS	Laurence	ICPAL
Mme	GRAVE	Béatrice	Toxicologie
Mme	GROSS	Barbara	Biochimie
M.	HAMONIER	Julien	Biomathématiques
Mme	HAMOUDI	Chérifa Mounira	Pharmacotechnie industrielle
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie
Mme	HELLEBOID	Audrey	Physiologie
M.	HERMANN	Emmanuel	Immunologie
M.	KAMBIA	Kpakpaga Nicolas	Pharmacologie
M.	KARROUT	Youness	Pharmacotechnie Industrielle
Mme	LALLOYER	Fanny	Biochimie
M.	LEBEGUE	Nicolas	Onco et Neurochimie
Mme	LECOEUR	Marie	Chimie Analytique
Mme	LEHMANN	Hélène	Législation
Mme	LELEU-CHAVAIN	Natascha	ICPAL
Mme	LIPKA	Emmanuelle	Chimie Analytique
Mme	MARTIN	Françoise	Physiologie
M.	MOREAU	Pierre Arthur	Sciences végétales et fongiques
M.	MORGENROTH	Thomas	Législation
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle
Mme	NIKASINOVIC	Lydia	Toxicologie
Mme	PINÇON	Claire	Biomathématiques
M.	PIVA	Frank	Biochimie
Mme	PLATEL	Anne	Toxicologie
M.	POURCET	Benoît	Biochimie
M.	RAVAUX	Pierre	Biomathématiques
Mme	RAVEZ	Séverine	Onco et Neurochimie
Mme	RIVIERE	Céline	Pharmacognosie
Mme	ROGER	Nadine	Immunologie
M.	ROUMY	Vincent	Pharmacognosie
Mme	SEBTI	Yasmine	Biochimie

Mme	SINGER	Elisabeth	Bactériologie
Mme	STANDAERT	Annie	Parasitologie
M.	TAGZIRT	Madjid	Hématologie
M.	VILLEMAGNE	Baptiste	Lab. de Médicaments et Molécules
M.	WELTI	Stéphane	Sciences végétales et fongiques
M.	YOUS	Saïd	Onco et Neurochimie
M.	ZITOUNI	Djamel	Biomathématiques

Professeurs Certifiés

Civ.	NOM	Prénom	Laboratoire
M.	HUGES	Dominique	Anglais
Mlle	FAUQUANT	Soline	Anglais
M.	OSTYN	Gaël	Anglais

Professeur Associé - mi-temps

Civ.	NOM	Prénom	Laboratoire
M.	DAO PHAN	Hai Pascal	Lab. Médicaments et Molécules
M.	DHANANI	Alban	Droit et Economie Pharmaceutique

Maîtres de Conférences ASSOCIES - mi-temps

Civ.	NOM	Prénom	Laboratoire
M.	BRICOTEAU	Didier	Biomathématiques
Mme	CUCCHI	Malgorzata	Biomathématiques
M.	FRIMAT	Bruno	Pharmacie Clinique
M.	GILLOT	François	Droit et Economie pharmaceutique
M.	MASCAUT	Daniel	Pharmacie Clinique
M.	ZANETTI	Sébastien	Biomathématiques
M.	BRICOTEAU	Didier	Biomathématiques

AHU

Civ.	NOM	Prénom	Laboratoire
Mme	DEMARET	Julie	Immunologie
Mme	HENRY	Héloïse	Biopharmacie
Mme	MASSE	Morgane	Biopharmacie

REMERCIEMENTS

Au Président de mon jury

Le Professeur Thierry Brousseau
Professeur des Universités – Praticien Hospitalier
Institut de Biochimie
Centre de Biologie et Pathologie
Centre Hospitalier Régional et Universitaire de Lille

Vous me faites l'honneur de présider ce jury de thèse.

Je vous remercie pour vos conseils et votre soutien tout au long de mon internat.

Le Professeur Marie-Pierre Buisine
Professeur des Universités – Praticien Hospitalier
Institut de Biochimie
Centre de Biologie et Pathologie
Centre Hospitalier Régional et Universitaire de Lille

Je vous remercie pour l'intérêt que vous portez à ce travail et pour avoir accepté de juger cette thèse.

Soyez assurée de mes sincères remerciements.

À mes directeurs de thèse

Le Docteur Julie Leclerc
Maitre de Conférence des Universités - Praticien Hospitalier
Institut de Biochimie
Centre de Biologie et Pathologie
Centre Hospitalier Régional et Universitaire de Lille

Merci à toi pour ton encadrement clair et complet, cette thèse fut pour moi l'occasion de travailler avec toi et ce fut un réel plaisir.

Le Docteur Jamal Ghoumid
Maitre de Conférence des Universités - Praticien Hospitalier
Institut de Génétique Médicale
Hôpital Jeanne de Flandres
Centre Hospitalier Régional et Universitaire de Lille

Merci pour tes conseils et le partage de ta vision clinique, ce type de collaboration est aujourd'hui essentielle à notre activité clinico-biologique.

Je tiens également à présenter mes remerciements :

A tous les biologistes qui ont partagé leur expérience quotidienne en biologie moléculaire pour la rédaction de ce mémoire : Thomas Smol, Adrien Pagin, Claire-Marie Dhaenens, Frank Broly, Nicole Porchet, Christophe Zawadzki, Isabelle Fajardy, Fabienne Escande, Isabelle Vuillaume, Monique Fontaine, Joseph Vamecq, Catherine Vermaut, Claude Preudhomme et Nicolas Duployez.

A toute l'équipe de bioinformatique Lilloise pour leur aide à la rédaction de ce mémoire, mais aussi pour leur aide quotidienne : Emilie, Christophe, Fabrice.

Aux Docteurs Malika Balduyck et Marie-Françoise Odou, pour le partage de leur expérience de biologiste moléculaire qui m'a permis la rédaction de ce mémoire, mais également pour tous leurs conseils concernant l'enseignement.

A Anne Sophie Jourdain et Caroline Thuillier pour leur aide, la découverte de la génétique mais aussi pour leur sympathie au quotidien.

A Benjamin Lopez, pour son aide plus que précieuse en statistiques.

A tous mes co-internes de biologie et clinique qui ont fait que l'internat fut aussi le moment de très belles rencontres amicales : Gauthier, Zoé, Alexandre, Stéphanie, Florian, Damien, Réna, Kada, Valérie, Maxime, Maud, David, Martin, Angélique, Elise, Frédérique, Caroline, Nicolas, Jean-David, Olivier, Adélaïde, Marie, Méline, Pauline, Jennifer, mais également la relève Estelle, Camille, Claire, Doriane, Léo-Paul, Youssef.

A mes co-internes de génétique qui ont partagé avec moi le temps d'un DU la passion de la bioinformatique : Elouanh, Mathilde, Pierre.

A mes co-équipiers de l'association des internes qui m'ont accompagné durant 4 mandats : Elodie(s), Marie, Julie, Samy, Terry, Aurélien, Anthony, Sarah, Mathieu, Guillaume, Samir et Héloïse.

A tous les biologistes et cliniciens qui m'ont accompagné pendant mon cursus.

A tous les techniciens et ingénieurs qui m'ont formé et accompagné au cours de mon cursus d'interne, pour leur accueil et leur sympathie.

A ma famille, qui a accepté et compris les enjeux d'un métier si passionnant.

A ma femme Amandine, qui m'accompagne au quotidien, me soutient et accepte tous mes choix professionnels.

SOMMAIRE

Résumé

L'utilisation du séquençage de nouvelle génération tend à se généraliser dans les laboratoires de génétique moléculaire. En effet, cette technologie permet l'analyse rapide d'un grand nombre de gènes chez un grand nombre de patients, dans des délais relativement courts. Toutefois, la contrepartie est l'identification d'un très grand nombre de variants par gène étudié et par patient. A l'échelle d'un génome entier, plus de 3 millions de variations sont retrouvées. Cependant la grande majorité n'auront pas d'effet délétère. Il convient donc d'avoir une méthode robuste de caractérisation des variants afin de déterminer lequel ou lesquels sont responsables de la pathologie. Une des étapes clés du traitement bioinformatique des données est « l'annotation » des variants, qui permet de décrire le variant selon plusieurs critères, incluant sa fréquence dans la population générale, sa présence dans les bases de données ou encore les prédictions *in silico* quant à leur pathogénicité. En prenant en compte tous ces éléments et le phénotype clinique du patient, le biologiste interprète et classe les variants identifiés selon des recommandations telles que celles de *l'American College of Medical Genetics and Genomics* (ACMG), en cinq classes différentes, allant de « variant pathogène» à «variant bénin»).

Dans certains cas, il n'y a pas suffisamment de preuves pour déterminer si les variants identifiés sont responsables de la pathologie. L'utilisation d'outils de prédiction *in silico* peut aider le biologiste dans l'interprétation. Cependant, ces outils ne suffisent pas à eux seuls et sont parfois peu informatifs. Le variant est alors classé en variant « de signification inconnue », et ce malgré les réunions de concertations pluridisciplinaires (RCP). Afin de sortir de cette impasse, le recours à des analyses fonctionnelles est nécessaire pour mieux comprendre l'impact du variant sur l'expression ou la fonction de la protéine, et donc son implication dans le phénotype du patient. Nous souhaitons faire un état des lieux des pratiques et des stratégies utilisées au sein du CHRU de Lille, d'une part vis-à-vis de l'utilisation des outils *in silico* dans l'interprétation de ces variants, mais également vis-à-vis des tests fonctionnels afin de savoir dans quelles mesures ceux-ci peuvent être intégrés dans le processus de diagnostic. L'objectif est de déterminer d'une part la performance des outils *in silico* utilisés, et d'autre part de donner quelques exemples de tests fonctionnels qui aident à la classification des variants identifiés en NGS.

I. Le séquençage

1. Historique du séquençage

Les techniques de séquençage sont nées dans les années 1970 et ont valu un prix Nobel à leurs inventeurs : Maxam et Gilbert, et Sanger. Ces techniques ont évolué au fil des données et sont devenues de plus en plus performantes, jusqu'à permettre aujourd'hui de séquencer un génome entier en un temps court.

1.1. Méthode de Maxam et Gilbert

Il s'agit d'une méthode chimique qui repose sur la coupure de la molécule d'ADN marquée radioactivement par le ^{32}P. C'est donc une méthode par dégradation. Le clivage est réalisé grâce à des réactifs chimiques à l'extrémité 5' ou 3' au niveau d'une base spécifique à cette réaction. Les fragments obtenus sont séparés par migration sur gel. La longueur des fragments marqués permet l'identification de la position de la base impliquée dans la coupure et le réactif chimique responsable de la coupure permet l'identification de la base impliquée (Figure 1). Cette technique n'a pas eu de développement ultérieur majeur en raison de la toxicité des réactifs chimiques utilisées, de la difficulté d'automatisation ainsi que de la taille limitée des fragments séquencés (<250 pb) (1).

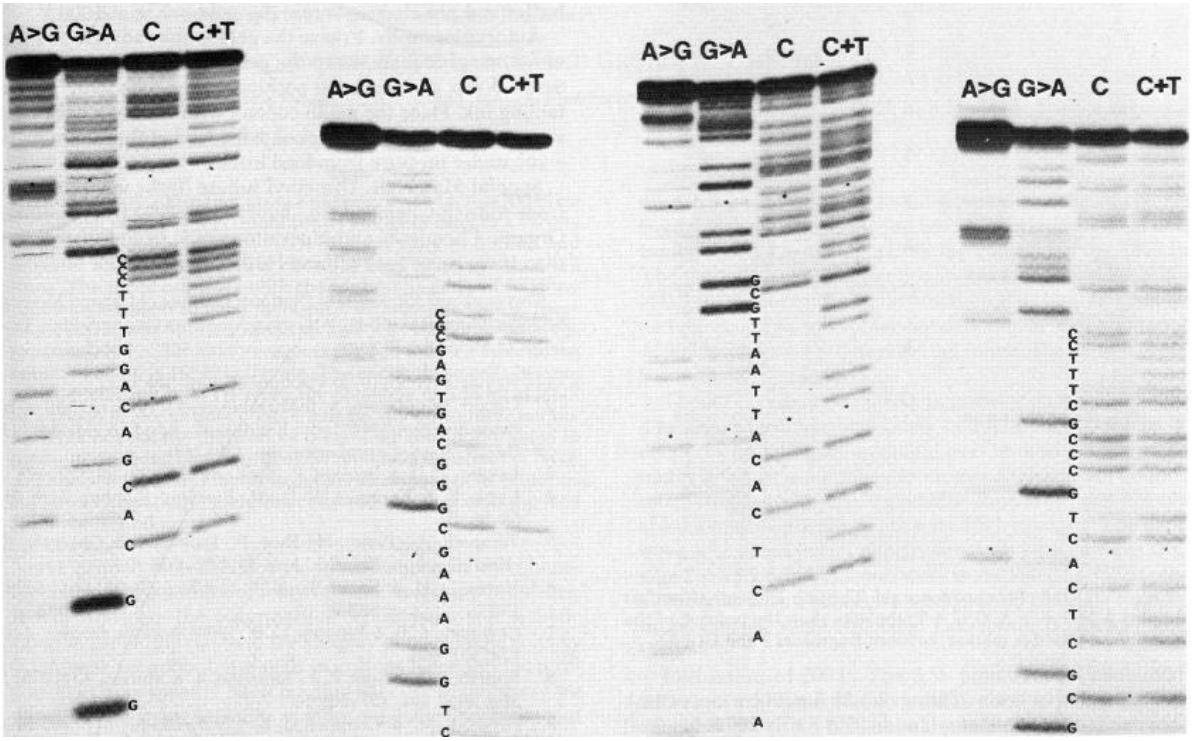

Figure 1. Autoradiogrammes de gels de séquence après séparation électrophorétique et lecture du profil des bandes radioactives selon la méthode de Maxam et Gilbert. (1).

1.2. Méthode de Sanger

Cette technique est basée sur la synthèse enzymatique de l'ADN en présence de nucléotides modifiés radiomarqués qui vont arrêter l'élongation (méthode par synthèse). Les nucléotides modifiés, les didésoxyribonucléotides (ddNTP), sont dépourvus du groupement hydroxyl situé en 3' du ribose (Figure 2). La conséquence de cette modification est l'impossibilité d'établir la liaison phosphodiester entre ce ribose modifié et le groupement phosphate du nucléotide suivant, entrainant alors un arrêt de la synthèse du brin d'ADN. La réaction nécessite le mélange de nucléotides modifiés en faible quantité et de nucléotides normaux de façon à ce qu'il y ait incorporation à la fois d'un dNTP et d'un ddNTP à chaque position. Des fragments de tailles variables sont alors obtenus et séparés sur un gel de polyacrylamide, et la longueur du fragment associé au marquage du nucléotide permet la lecture de la séquence (Figure 3) (2).

Cette méthode présente l'avantage d'être moins toxique et plus rapide que celle de Maxam et Gilbert et permet le séquençage de fragments plus longs (jusque 700 pb). Cette méthode a permis de séquencer le premier organisme entier, le bactériophage phiX174 (5375 pb) (3).

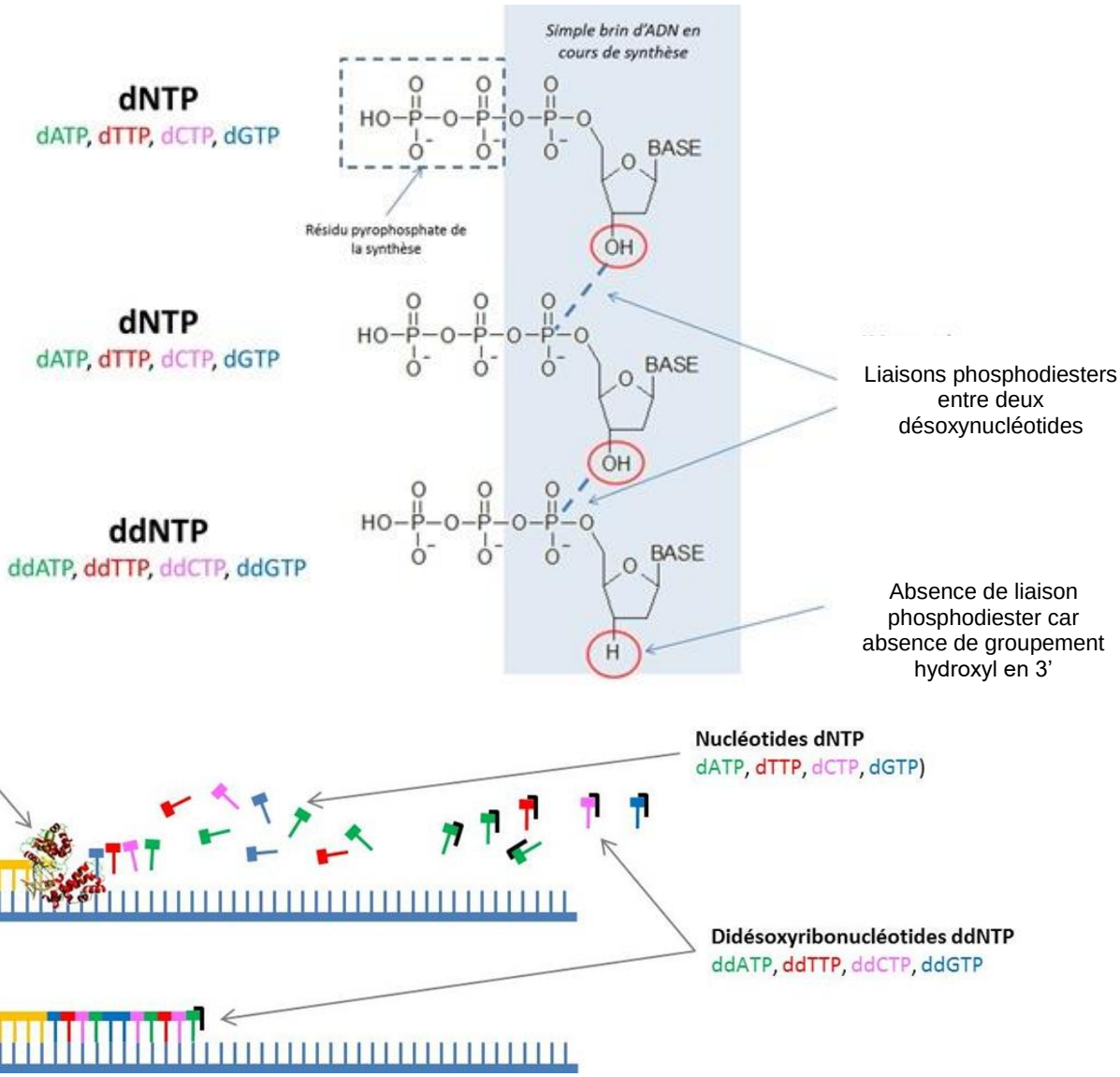

Figure 2. Principe du séquençage Sanger par utilisation de DiDéoxynucléotide (ddNTP) et de Déoxynucléotide (dNTP) (4).

Figure 3. Autoradiogrammes de gels de séquence après séparation électrophorétique et lecture du profil des bandes radioactives (2).

1.3. Automatisation du séquençage (séquençage de 1ère génération)

La technique de Sanger a progressivement été automatisée. Pour cela, plusieurs modifications ont été apportées. Premièrement, le marquage n'est plus réalisé grâce à la radioactivité mais par des nucléotides porteurs de fluorophores qui émettent en fonction du type de nucléotides à 4 longueurs d'ondes différentes. Chaque ddNTP émet donc un signal spécifique permettant de repérer individuellement chaque type de nucléotide. Secondairement, les fragments sont séparés par électrophorèse capillaire ; la miniaturisation du système permet de travailler simultanément sur un plus grand nombre d'échantillons. Les résultats sont obtenus par lecture successive des signaux de fluorescence à la sortie du capillaire, permettant après traitement informatique d'obtenir un électrophorégramme (Figure 4) (5).

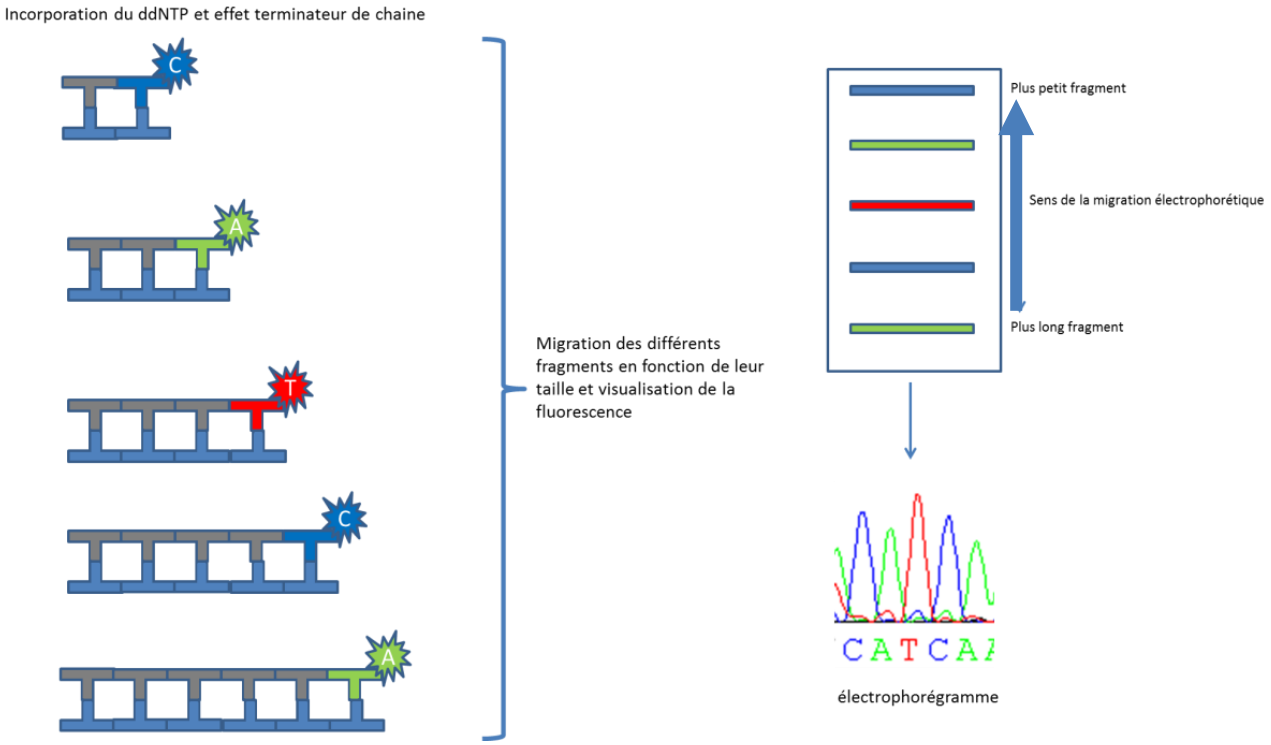

Figure 4. Principe de la méthode Sanger automatisée par utilisation de ddNTP fluorescents et de capillaires en verre.

1.4. Le projet Génome Humain

L'idée de séquencer le génome humain entier a été proposée par Renato Dulbecco en 1985 (prix Nobel en 1975 pour la découverte des oncogènes) (6). Après plusieurs années de discussion, le projet débute en 1989 et est prévu pour une durée de 15 ans avec un budget estimé à 3 milliards de dollars. En prélude au séquençage du génome humain, des

génomes d'autres espèces seront également séquencés. Ce projet est piloté par le *National Institutes of Health* (NIH) et dirigé par le Dr. James Watson (découverte de la conformation de l'ADN (7)). De nombreux centres de séquençage participent à ce projet d'ampleur internationale. En 1996, pour faciliter l'avancement du projet, le principe des Bermudes est adopté permettant le libre accès immédiat aux données de séquençage générées. Le projet est concurrencé en 1998 par une société privée (*Celera Genomic)* dirigée par Craig Venter. La course au séquençage se terminera alors en 2001, soit 3 ans avant la date initialement annoncée par le *NIH* (Figure 5) (8).

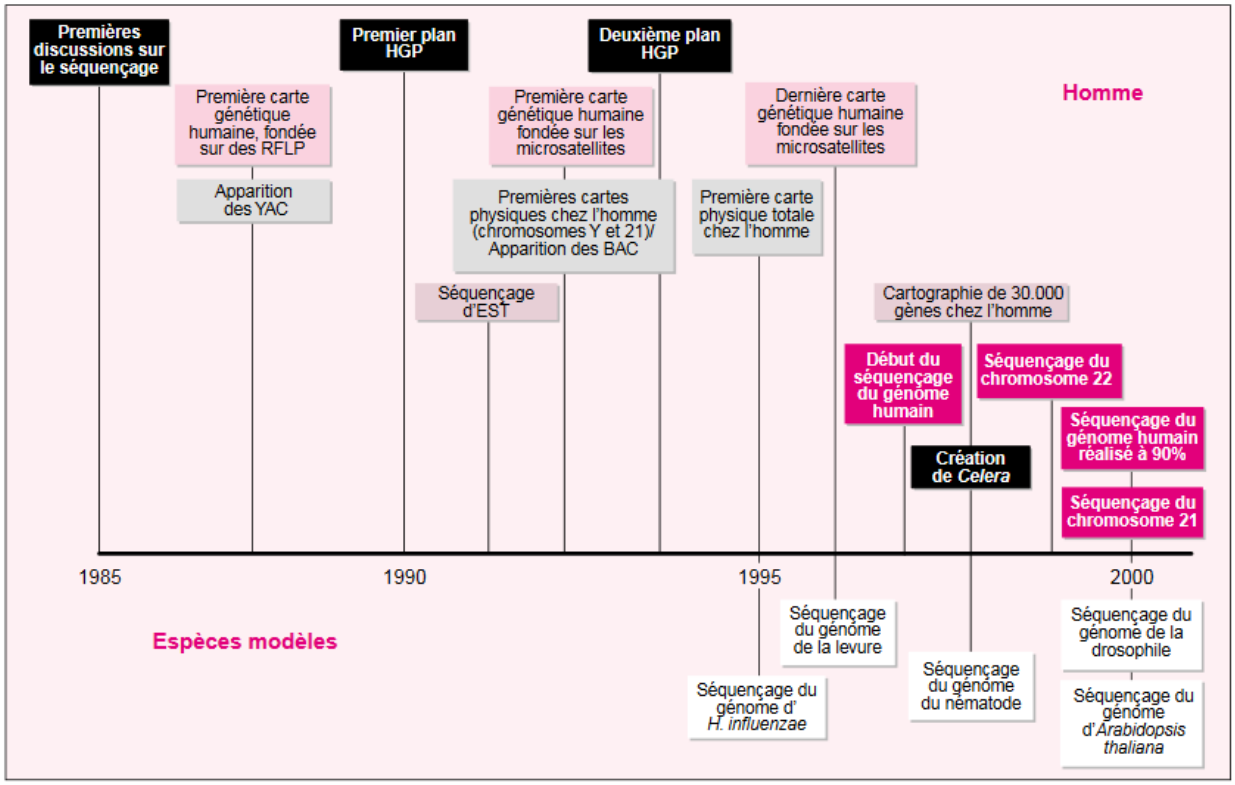

Figure 5. Principales étapes du projet génome humain (9).

Deux approches ont été utilisées dans le cadre de ce projet : une approche de séquençage dite de séquençage global aléatoire « *shotgun intégral*» et une approche par ordonnancement hiérarchique. Ces deux approches sont tributaires de l'informatique.

1.4.1. L'approche par séquençage global aléatoire

Cette approche commence par découper de façon aléatoire par sonication le génome en différentes sections : 2.000 paires de base, 10.000 paires de base ou 50.000 paires de base. Ces fragments sont ensuite amplifiés clonés dans des vecteurs amplifiés et séquencés. Des algorithmes mathématiques permettent ensuite d'assembler les

fragments qui se suivent (*contigs)* et de leur attribuer leur véritable emplacement dans le génome.

1.4.2. L'approche par ordonnancement hiérarchique

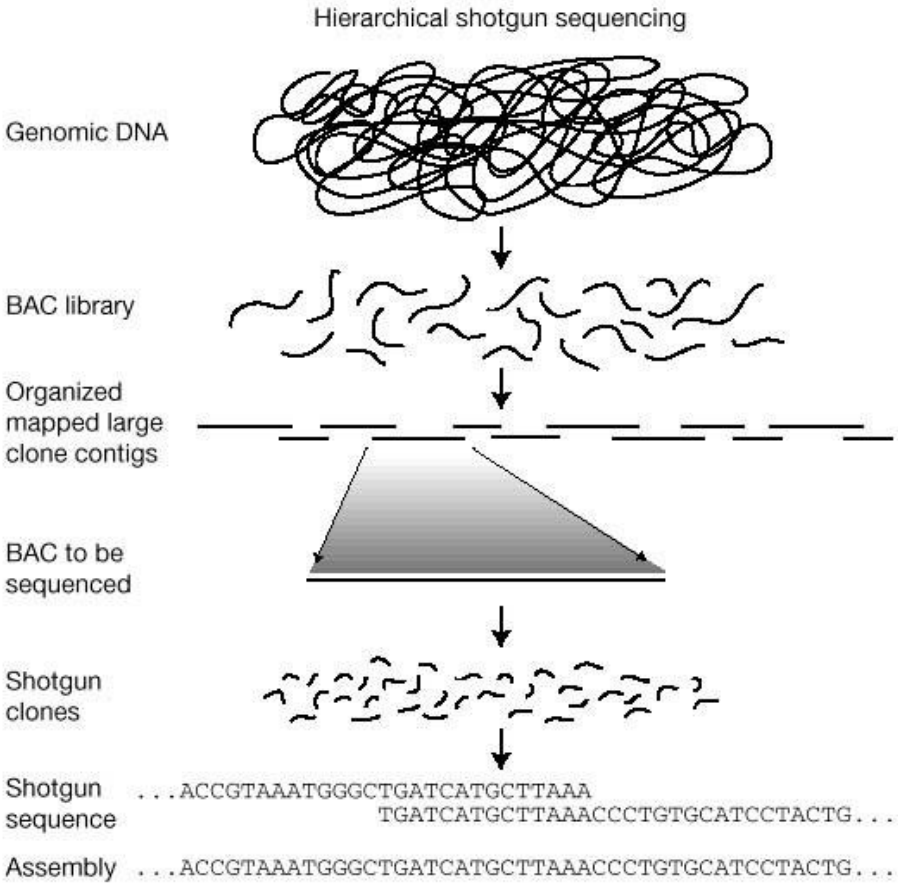

Figure 6. Représentation de la stratégie de séquençage par ordonnancement hiérarchique (8).

Le génome est découpé en fragments qui sont insérés dans un vecteur (chromosomes artificiels bactériens : BAC). Les BAC sont analysées pour obtenir une cartographie physique, il y a alors construction de contigs par comparaison de ces BACS. Ils sont sous-clonés en fragments de petite taille et ordonnés après sélection d'un nombre de clones dans chaque contig qui sont sous clonés par *shotgun* puis séquencés. Les séquences obtenues sont assemblées, ce qui permet une reconstitution des séquences de chaque BAC (Figure 6).

1.5. Vers le séquençage haut débit

En 2004, la première vague de développement est lancée dans le but de proposer un séquençage du génome pour le plus grand nombre et de baisser les coûts de façon à passer sous les 1000 $ pour le séquençage d'un génome humain (10). Dès 2001, les séquences sont obtenues à l'aide de chimies basées sur la méthode de Sanger et d'instruments capillaires (plateformes de séquençage «de première génération »). À partir de 2005, les plateformes de séquençage «de seconde génération » émergent et permettent une diminution importante des coûts (11). Aujourd'hui, les coûts de séquençage sont de plus en plus faibles ; les progrès de la science génomique, plus rapides que ne le prévoyait la loi de Moore, ont permis d'accélérer l'avènement du séquençage haut débit en diagnostic (Figure 7) (12).

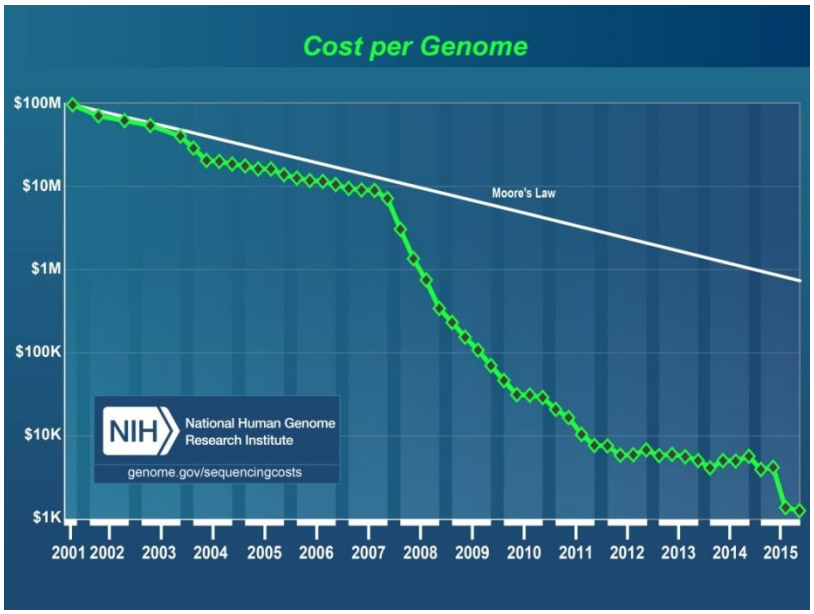

Figure 7. Evolution du coût des génomes (12).

2. Le séquençage haut débit (SHD ou NGS)

L'essor du diagnostic moléculaire par séquençage haut débit a été catalysé par le développement de nombreuses technologies de séquençage. La production peu coûteuse de gros volumes de données de séquence est l'avantage principal par rapport aux méthodes classiques (13). Les séquenceurs ont évolué, proposant des longueurs de fragments séquencés et une quantité de bases séquencées de plus en plus importantes, ainsi qu'une qualité croissante des séquences permettant une utilisation en diagnostic (Figure 8). Le développement de ces technologies permet aujourd'hui de pouvoir réaliser

le séquençage d'un génome de patient à un coût de séquençage aux alentours des 1000$.

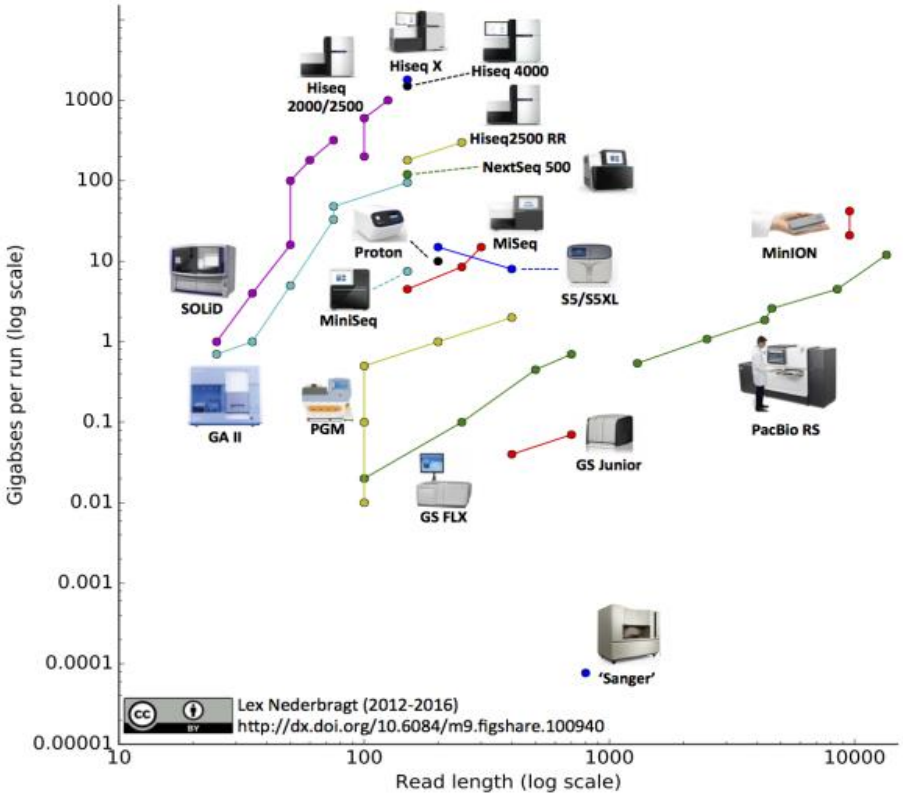

Figure 8. Evolution des séquenceurs en fonction de la longueur des séquences (reads) et de la quantité de données générées (14).

Cependant, l'analyse par séquençage haut débit ne se résume pas au séquenceur seul et nécessite tout un cheminement d'étape, à la fois en amont et en aval du séquençage.

2.1. Types d'analyses : panels, WES et WGS

Le séquençage à haut débit permet l'analyse simultanée de nombreux gènes. En fonction des besoins diagnostiques, il est donc possible d'analyser soit un nombre limité de gènes à travers des panels de plus ou moins grande taille, soit l'ensemble des régions codantes des gènes (exome ou *WES Whole Exome Sequencing*), soit l'ensemble du génome (WGS ou *Whole Genome Sequencing*) (15). Le coût et le temps d'analyse augmentent avec la complexité du séquençage, que ce soit au niveau de l'investissement matériel (séquenceur) ou des investissements analytiques et post-analytiques (informatique, stockage, personnel). L'augmentation du nombre de régions génomiques séquencées augmente la probabilité d'identifier des variants de signification inconnue, et également de réaliser des découvertes fortuites. Les découvertes fortuites correspondent à des découvertes d'altérations pathogènes sur un ou des gènes *a priori* sans rapport avec le

diagnostic clinique mais d'intérêt médical, et plus particulièrement sur les gènes dits « actionnables ». Ce sont des gènes pour lesquels il existe des recommandations spécifiques de prise en charge lorsqu'une altération pathogène est retrouvée (16). L'analyse par WES et WGS permet de réanalyser de façon exhaustive les données en cas de changement d'hypothèses diagnostiques ou de la littérature (Tableau 1). La complexité croissante des analyses permet donc d'augmenter le rendement diagnostique mais impose également des difficultés supplémentaires.

Analyse	**Indications**	**Exemples**	**Taille**	**Analyses**	**Variants de signification inconnus (VSI) et découvertes fortuites (DF)**	**Stratégie de réanalyse**
Panels Ciblés	Pathologies bien identifiées cliniquement et génétiquement.	Mucoviscidose (*CFTR*) Achondroplasie (*FGFR3*)	Quelques dizaines de kilobases (kb)	Analyses rapides, régions bien connues, bases de données bien documentées.	Peu de VSI, pas de DF	Non ou très limitée
Panels larges	Hétérogénéité génétique avec gènes identifiés et gènes candidats.	Panel épilepsie Panel cardiomyopathie	Plusieurs dizaines de kb	Temps d'analyse dépendant de la corrélation phénotype-génotype et de la bibliographie.	Quelques VSI, DF peu probables si panel restreint.	Possible mais limitée uniquement aux gènes déjà étudiés.
WES	Hétérogénéité génétique avec de nombreux gènes candidats	Déficience intellectuelle	200 000 exons 50 Mb	Temps d'analyse important, analyse difficile. Travail bibliographique important.	Nombreux VSI et DF probables.	Possible sur l'ensemble des gènes
WGS	Pathologies très hétérogènes. Anomalies moléculaires mal caractérisées.	Autisme	3,2 Gb	Analyse longue et fastidieuse, bibliographie peu documentée.	VSI très nombreux à la fois sur les régions exoniques, introniques et intergéniques. Autant de DF que pour le WES	Possible à tout niveau

Tableau 1. Comparaison des différentes stratégies de séquençage.

L'apport de l'exome dans le diagnostic moléculaire a notamment été montré dans un projet publié en 2014 dans le cadre des anomalies du développement (projet DDD : *Deciphering Developmental Disorders*). Mille cent treize exomes en trios de patients, sans diagnostic établi et recrutés par les 24 centres de génétique clinique du Royaume-Uni, ont été séquencés par le Wellcome Trust Sanger Institute. Le séquençage de trios (enfant/parents) a permis l'identification de l'anomalie génétique dans 27% des cas. Ce qui triple le rendement diagnostique par rapport aux analyses par panels de gènes (17).

Le séquençage du génome complet permet quant à lui une meilleure couverture de l'exome, donc une augmentation du rendement diagnostique. Ceci a notamment été montré par le séquençage de 50 trios de patients dans le cadre de la déficience intellectuelle sévère en identifiant la mutation dans 47% des cas (18).

2.2. La technique

La technique est séparée en 3 étapes : la préparation des échantillons (librairie), l'amplification clonale et le séquençage.

2.2.1. Préparation de la librairie

Une librairie est préparée à partir des ADN des patients. La librairie correspond à l'ensemble des régions ciblée : panel de gènes, WES ou WGS en fonction de ce que l'on souhaite séquence. C'est l'étape d'enrichissement. Cette étape permet de combiner plusieurs patients dans une même analyse de séquençage grâce aux index (ou codes-barres) qui sont des séquences nucléotidiques synthétiques ajoutées au fragment. Plusieurs approches permettent de construire des librairies : basées sur la constitution d'amplicons ou sur la capture (Figure 9).

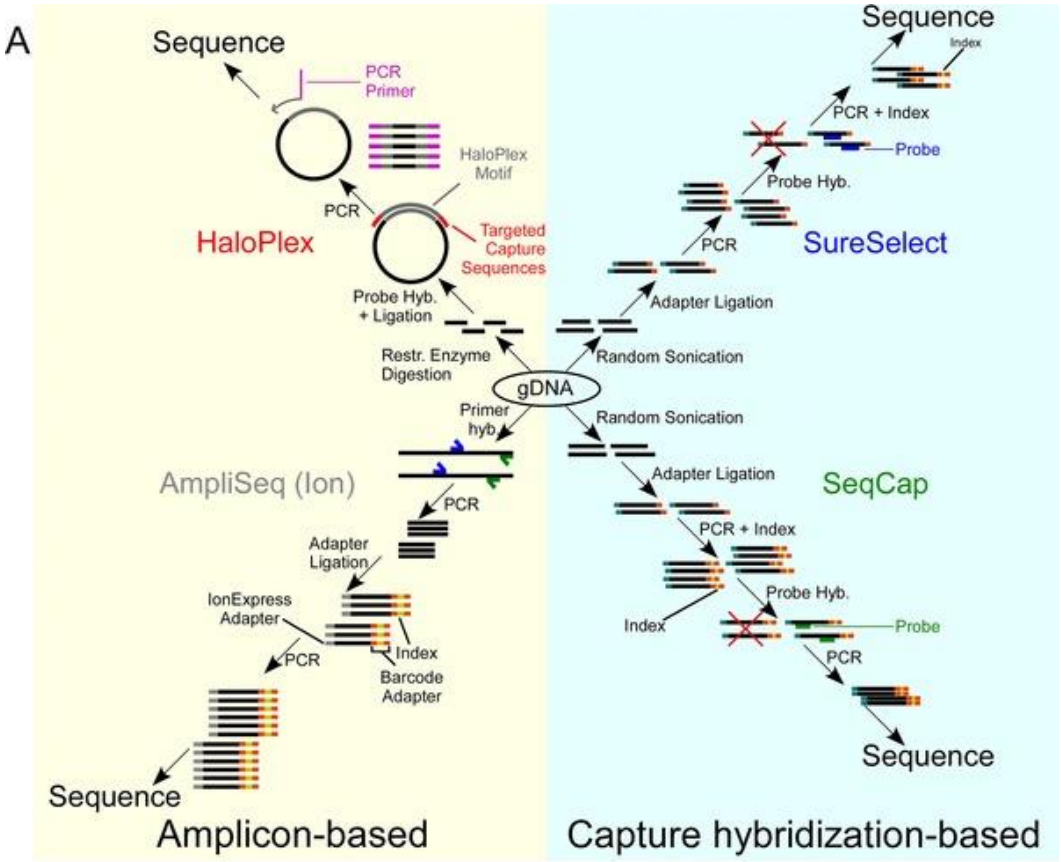

Figure 9. Méthodes d'enrichissement pré-séquençage (19).

2.2.1.1. Enrichissement par amplicons

Cette approche utilise la PCR Multiplex pour réaliser une amplification des régions d'intérêt. Il y a ensuite digestion des amorces et ligation des adaptateurs et barcodes (Figure 10). Cette approche est simple mais nécessite de bonnes quantité et qualité d'ADN de départ.

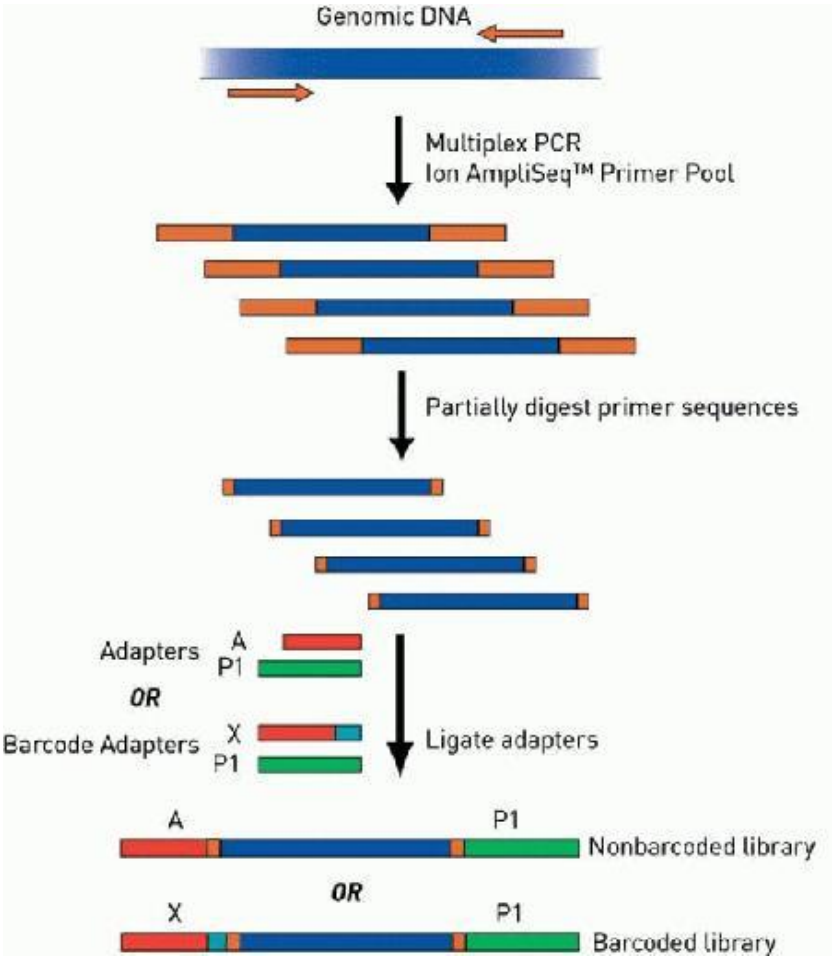

Figure 10. Enrichissement par obtention d'amplicons. Exemple de l'Ampliseq (20).

2.2.1.2. Enrichissement par capture

Dans cette stratégie, après fragmentation de l'ADN, des adaptateurs sont ajoutés par ligation. Il y a ensuite hybridation des séquences d'intérêt à des sondes puis un lavage de l'ADN est réalisé afin d'éliminer les régions non hybridées grâce à des billes magnétiques ou de streptavidines (Figure 11).

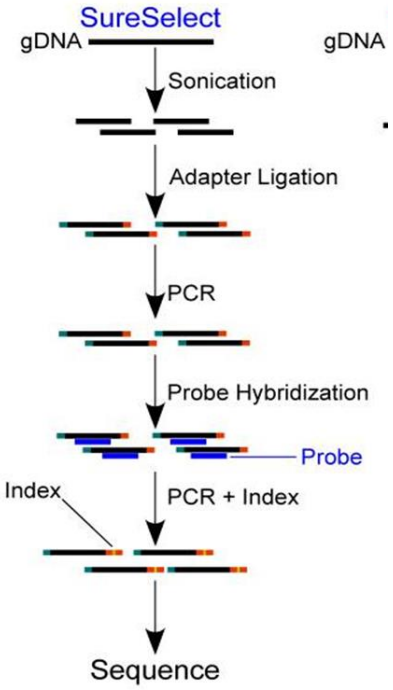

Figure 11. Exemple d'enrichissement par capture (21).

2.2.1.3. Enrichissement par circularisation

Cette approche utilise une sonde partiellement double brin dont les extrémités simples brins sont complémentaires des extrémités de la région d'intérêt. L'ADN est fragmenté par un cocktail d'enzymes de restriction. Les sites de coupure peuvent être prédits par un algorithme informatique et les extrémités des sondes peuvent être synthétisées pour être complémentaires des extrémités de la région d'intérêt. L'hybridation des sondes entraine une circularisation du fragment d'ADN qui est ensuite amplifié par PCR (Figure 12).

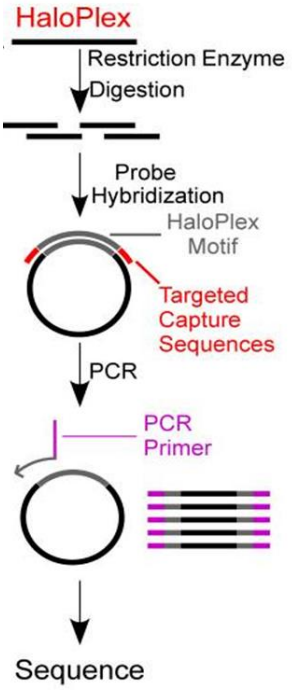

Figure 12. Exemple d'enrichissement par circularisation (21).

2.2.2. Amplification clonale

L'étape d'amplification clonale permet d'amplifier les séquences que l'on souhaite analyser, ceci afin d'obtenir un signal de séquençage détectable.

2.2.2.1. Bridge PCR

L'amplification est réalisée sur une surface solide (*flowcell*). Chaque fragment se lie à cette surface via son adaptateur. Les fragments sont amplifiés par PCR avec formation de ponts grâce aux séquences adaptatrices. Il y a alors obtention de groupes d'amplicons issus d'un même fragment initial (clusters) à la surface de la *flowcell* (Figure 13).

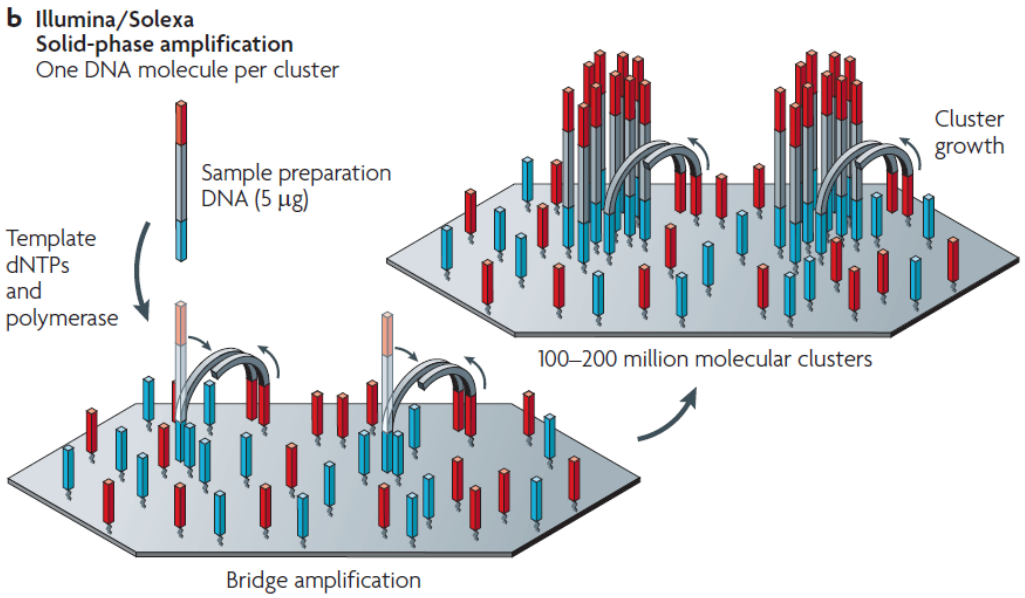

Figure 13. Etapes de l'amplification clonale par Bridge PCR (13).

2.2.2.2. PCR Emulsion

Grâce à la formation d'une émulsion (mélange huile/eau), un fragment d'ADN et une bille sont capturés au sein d'une goutte. Le fragment se lie à la bille via son adaptateur. Après plusieurs étapes d'amplification par PCR, il y a obtention de nombreux amplicons à la surface de la bille. Les billes avec fragments sont sélectionnées par formation d'un complexe biotine (sur l'adaptateur) streptavidine (sur des billes spécifiques) (Figure 14).

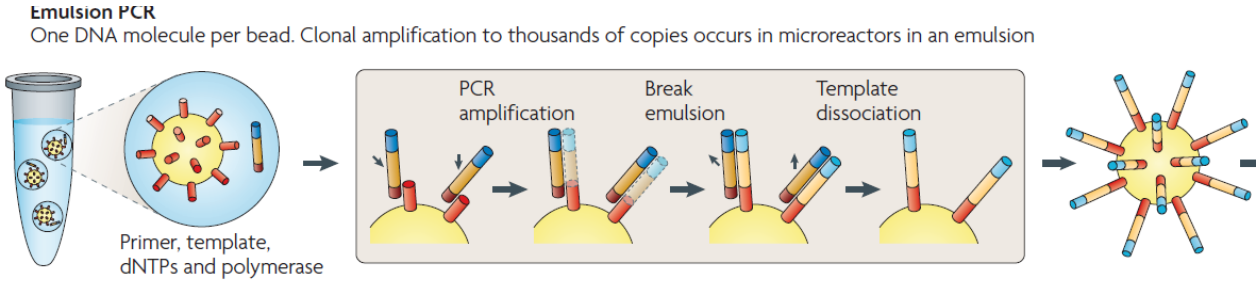

Figure 14. Etapes de l'amplification clonale par PCR émulsion (13).

2.2.3. Séquençage

2.2.3.1. Méthode Life (détection de protons)

Le séquençage « Ion Torrent® » est basée sur la libération de protons lors de l'ajout d'un nucléotide. Cette technique débute par l'hybridation d'une amorce universelle et par utilisation de nucléotides non marqués. Pour chaque cycle de séquençage, il y a injection séquentielle de chacune des bases (séparément) et mesure du pH. Si la base est intégrée

dans le fragment d'ADN en cours de synthèse, il y a libération de proton et variation du pH. L'intégration de plusieurs bases identiques lors d'une même injection est possible si la séquence est un homopolymère. Les modifications de pH sont ensuite converties en signal pouvant être lu et analysé (Figure 15).

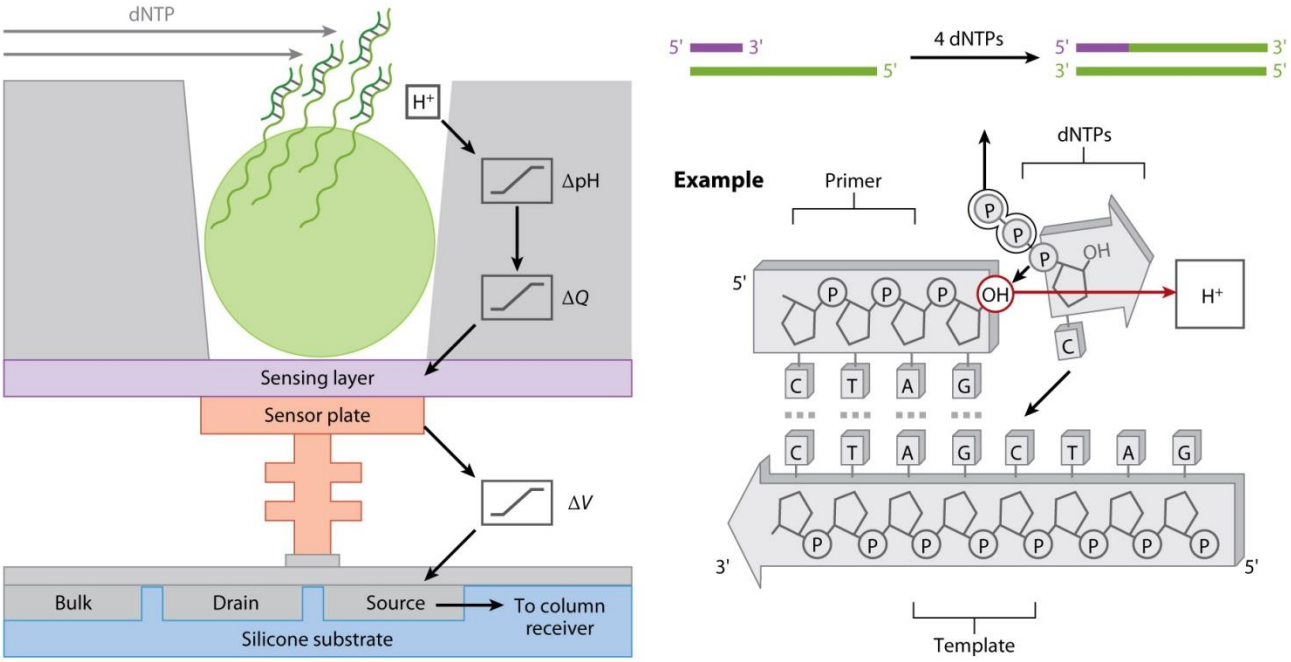

Figure 15. Séquençage basé sur la méthode IonTorrent®, mesure du pH après injection séquentielle de chaque base (22).

2.2.3.2. Méthode Illumina (détection de photons)

Le principe du séquençage « Illumina » est équivalent au séquençage Sanger : utilisation de nucléotides modifiés marqués par fluorescence, mais avec un terminateur de chaine réversible. Il débute par hybridation d'une amorce universelle. Pour chaque cycle de séquençage, il y a présentation des 4 bases simultanément et incorporation d'une base et une seule dans le fragment d'ADN en cours de synthèse avec émission d'une fluorescence et acquisition du signal lumineux (Figure 16).

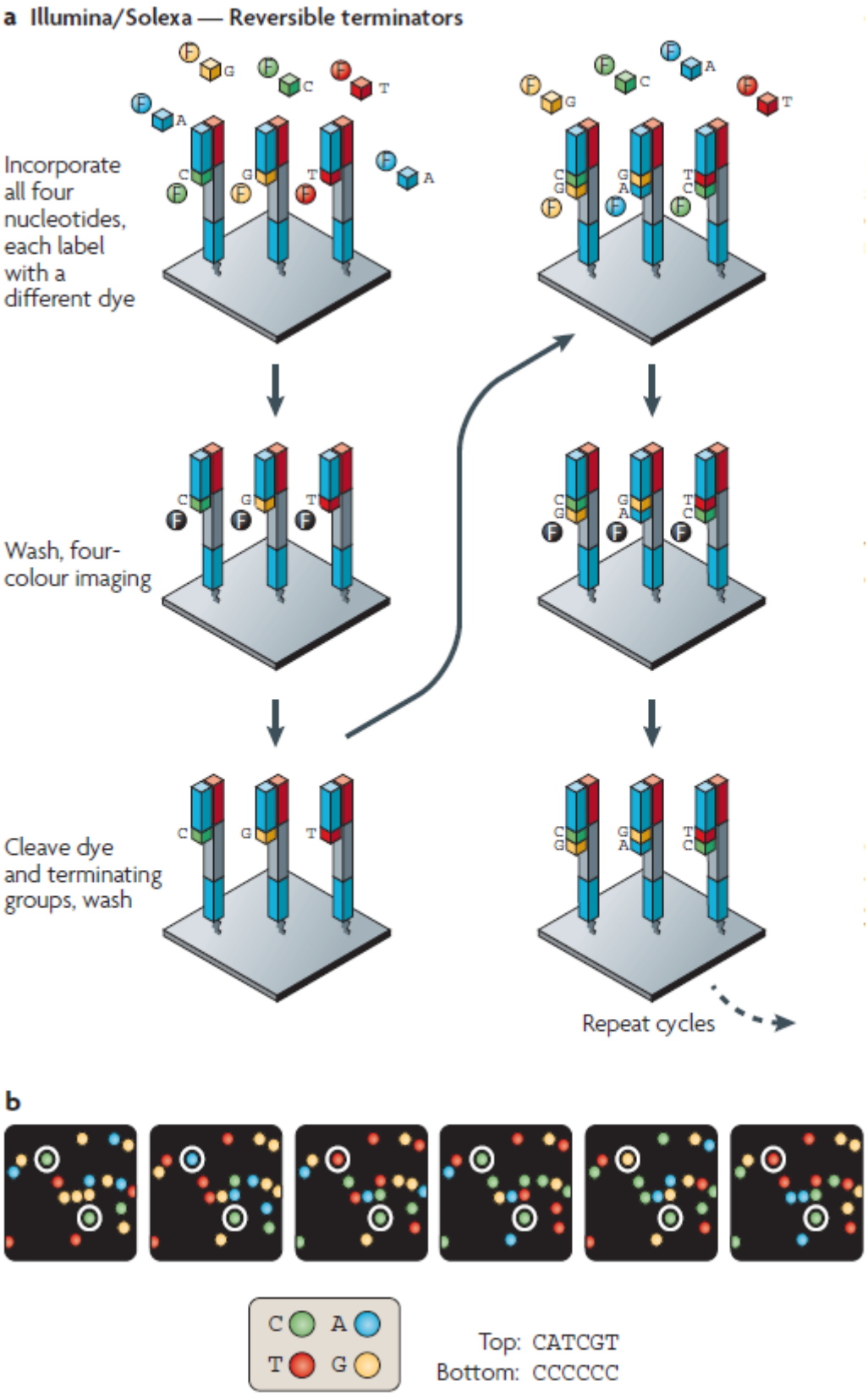

Figure 16. Séquençage par terminateurs réversibles (a) et signal fluorescent obtenu (b) (13).

2.3. Analyse bioinformatique des données

Les avancées en termes de séquençage ont apporté de nouvelles difficultés liées au traitement des données générées, qui sont massives et complexes. Ceci nécessite l'utilisation d'un nouveau domaine de compétence, la bioinformatique.

2.3.1. Architecture matérielle (Hardware)

Les principaux composants d'une architecture de séquençage haut débit sont les séquenceurs, les calculateurs hautes performances (HPC), les serveurs de stockage et le réseau (Figure 17). Les choix des différents éléments sont fonctions de la taille des séquences analysées. Par exemple, l'analyse d'un panel de dix gènes et celui d'un exome entier ne nécessiteront ni la même puissance de calcul, ni la même capacité de stockage.

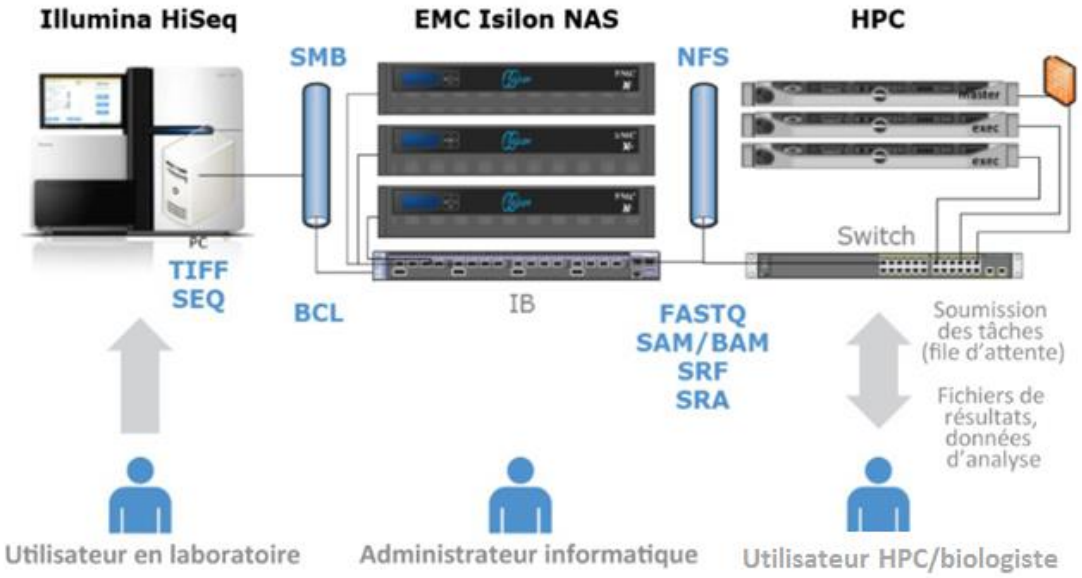

Figure 17. Exemple d'architecture informatique de référence comprenant un séquenceur (Illumina HiSeq), un serveur de stockage (EMC Isilon NAS), un calculateur (HPC) et un système réseau (switches, SMB et NFS) (23).

2.3.1.1. Réseau et serveur

Les données générées sous forme de fichiers par le SHD ne peuvent être analysées directement sur le séquenceur ; il faut les transférer sur un serveur de stockage externe. Ce serveur de stockage externe dit « rapide » est essentiel pour une analyse efficace des données par les serveurs de calcul. Il est souvent couplé à un serveur de stockage complémentaire dit « lent » qui permet d'archiver sur le long terme les fichiers et les résultats d'analyse. Les fichiers transitent via le réseau grâce à des protocoles de transfert et de partage de fichiers. Les protocoles les plus utilisés sont FTP (*File Transfer Protocol*),

Rsync (Remote Synchronization), SMB *(Server Message Block),* et NFS *(Network File System)*. Les données générées par le SHD sont très volumineuses et nécessitent un réseau qui soit à la fois rapide, fiable et sécurisé. L'utilisation de contrôles sur l'intégralité des données lors des transferts est préconisée (checksum par exemple).

2.3.1.2. Le serveur de stockage

L'unité de base informatique est le bit (*binary digit*), qui est un chiffre binaire (valeur 1 ou 0). L'unité de stockage de base est l'octet (1 o) qui correspond à 8 bits. En ordre de grandeur, un fichier texte fait environ 500 ko, une photo 10 Mo et un film 750 Mo. Un *run* de NextSeq 550 Illumina® représente environ 30 milliards de pb et les données brutes représentent par *run* environ 50 Go pour les images générées par le séquenceur, 50 Go pour les fichiers FASTQ et 50 Go pour les autres données (BAM, VCF …), soient 150 Go par *run*. Etant donnés les coûts importants de séquençage, la taille importante des fichiers et le besoin de conserver les données sur une longue période et de réaliser des sauvegardes supplémentaires *(backup)* des données générées. Les systèmes de redondance (à l'image de la technologie RAID *Redundant Array of Independant Disks*) des systèmes de stockage utilisées dans le SHD, permettent d'assurer la disponibilité des fichiers en cas de panne partielle des disques durs, ce qui est essentiel dans l'activité de séquençage. Le volume, en constante augmentation de la taille des fichiers nécessite un choix judicieux du type de fichier à conserver afin d'éviter la saturation rapide des serveurs. Ce stockage peut se réaliser en local sur des serveurs hébergés en interne ou de façon délocalisé grâce à des serveurs distants type Cloud mais ces serveurs doivent être certifiés « hébergeur de santé » car nous sommes dans le cadre de données médicales.

2.3.1.3. Le calculateur

Afin de traiter les données brutes et au vu de leur complexité et volume importants, il est nécessaire d'utiliser de puissantes machines de calculs. La puissance d'un calculateur est mesurée grâce au FPU *(floating point operation per second unit),* qui correspond à une opération élémentaire à virgule flottante (méthode d'écriture des nombres réels via un triplet : signe, mantisse et exposant). Ces opérations sont réalisables grâce aux processeurs *(CPU : Central Processing Unit)* multicœurs qui permettent de les réaliser en parallèle, permettant de réduire considérablement le temps d'analyse. Il est également nécessaire de disposer d'un volume important de mémoire vive ou RAM.

2.3.1.4. Matériel et coûts

Le coût théorique d'un stockage redondé performant est d'environ 1000 €/To. Si l'on prend l'exemple d'un NextSeq, et que l'on considère 8 runs par mois, le stockage des données représente environ 15.6 To/an en gardant l'ensemble des données, soit un cout de 15600 euros. En ne conservant que les fichiers BAM, VCF et FASTQ, cela ne représente plus qu'un stockage nécessaire de 5.7 To/an soit 5700 euros. La vitesse du réseau doit être au minimum de 1Gb/seconde pour assurer un transfert des données primaires en moins d'une heure ; une telle infrastructure réseau coûte environ 100 000€ à la création. Le serveur de calcul nécessitera plusieurs CPU ainsi que de la RAM proportionnelle au nombre de cœurs avec une fréquence importante, ce qui représente environ un investissement de 20 000 € par séquenceur. Ces différents coûts n'incluent ni le personnel, ni les maintenances, ni l'installation du matériel, ni le coût des locaux. À la vue des coûts importants, il est donc d'usage d'utiliser des *clusters* de calcul regroupant les ressources informatiques afin de les mutualiser, les rendre plus disponibles, en faciliter la gestion et en augmenter les capacités globales.

2.3.2. Etapes et outils des processus analytiques (Software)

Le processus bioinformatique d'analyse des données du séquençage haut débit se découpe en trois principales étapes : l'analyse primaire qui génère les séquences, l'analyse secondaire va de l'alignement jusqu'à la génération des VCF, donc identification des variations par rapport à la séquence de référence et l'analyse tertiaire correspond à l'interprétation des résultats (Figure 18).

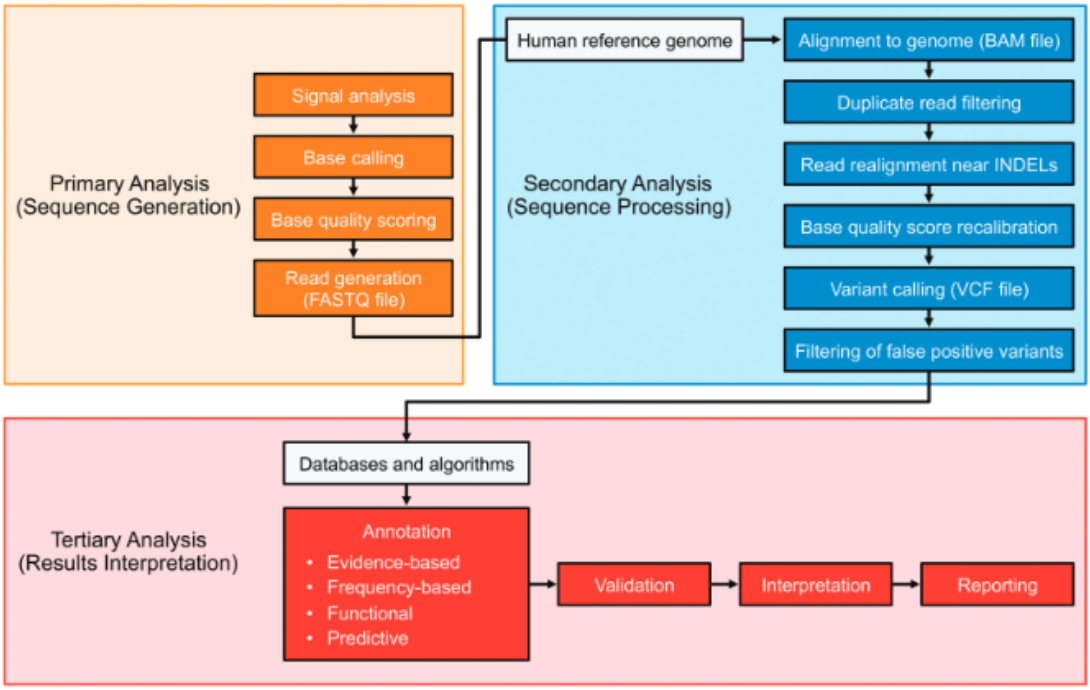

Figure 18. Principales étapes d'un *pipeline* bioinformatique utilisé pour l'analyse des données de séquençage à haut débit (24).

2.3.2.1. Analyse primaire

2.3.2.1.1. *Base Calling*

Le *Base Calling* permet de convertir le signal brut du séquenceur (lumière ou variation de pH) en séquence et également d'estimer la qualité des bases obtenues. Un fichier FASTQ (FASTA Quality) est alors généré (25) ; il est composé d'un bloc de trois lignes pour chaque séquence : l'identifiant de la séquence, la séquence nucléotidique et le score de qualité.

2.3.2.1.2. Demultiplexage

Grâce aux index des adaptateurs, le démultiplexage permet d'attribuer chaque séquence lue à un échantillon donné.

2.3.2.2. Analyse secondaire

2.3.2.2.1. *Trimming*

Il s'agit d'éliminer les séquences d'amorce et les bases de mauvaise qualité. La qualité est obtenue par la scorification de QPhred (26). Il s'agit de la probabilité p d'erreur de la base (QPhred = $-10\log_{10}p$). Par exemple Q20 correspond à une probabilité d'erreur de 1% et Q30 de 0.1%.

2.3.2.2.2. Alignement

Suite au *trimming*, les séquences obtenues sont alignées sur une séquence de référence. En fonction du type d'analyse, il est possible de réaliser cet alignement soit sur des séquences données ciblées, soit sur la totalité du génome. Plusieurs outils permettent de réaliser l'alignement, tels que Bowtie, BWA ou SAMtools (24,27). Le type de fichier résultant de l'alignement est le SAM *(Sequence Alignment Map)*, cependant il est plutôt d'usage d'utiliser le format compressé du SAM qui est le BAM *(Binary Alignment Map)*. Le fichier SAM se compose de onze composantes dont le terme CIGAR *(Concise Idiosyncratic Gapped Alignment Report),* qui est une codification des alignements par rapport à la séquence de référence (Figure 19) (28). Il est possible d'estimer la qualité des données du fichier SAM obtenu, les taux d'alignements, la distribution des scores d'alignement (de 70 à 95% attendus) et la profondeur sur les cibles.

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Figure 19. Exemple de fichier SAM, description des onze composantes du SAM (correspondant aux colonnes du fichier SAM) et description du code CIGAR (28).

Le fichier BAM peut être lu directement avec des outils informatiques tel que IGV, ce qui permet de visualiser à la fois la séquence des *reads,* leur sens, ainsi que la couverture et la profondeur (29, 30). La longueur de lecture correspond au nombre de bases

successives obtenues par une lecture donnée. La couverture est le nombre de bases séquencées par rapport au nombre de bases de la séquence ciblée initialement. La profondeur correspond au nombre de lectures à une position donnée. Elle est quantifiée par le terme X, par exemple 20X correspond à une profondeur de 20 lectures à une position donnée. IGV permet également de visualiser les différences par rapport à la séquence de référence (variants) (Figure 20).

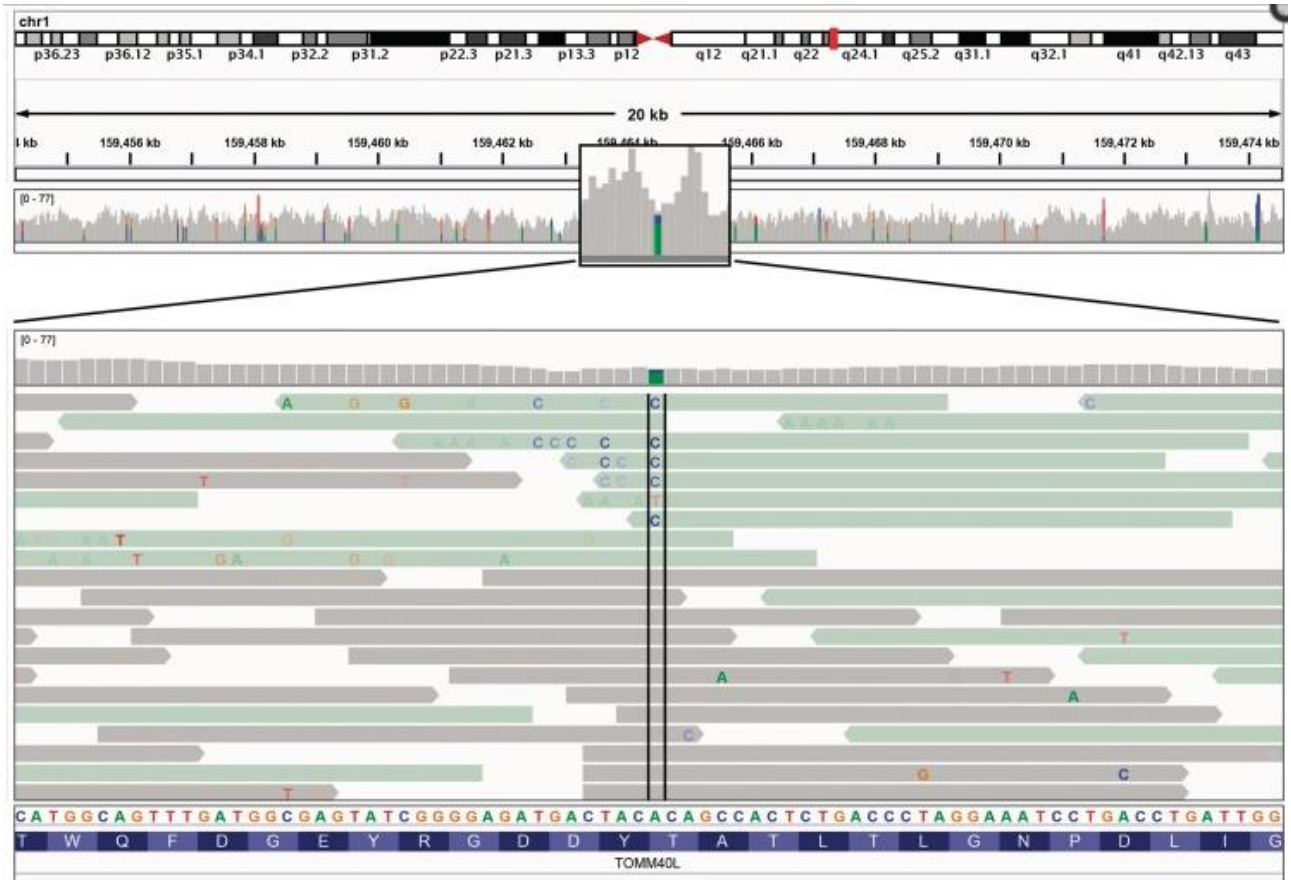

Figure 20. Visualisation d'un fichier BAM avec le logiciel IGV (Integrative Genome Viewer). Le sens des *reads* obtenu est montré par une lecture verte ou grise. Les variartions par rapport à la séquence de référence sont directement visualisés dans les lectures par le nucléotide variant (29).

2.3.2.2.3. Elimination des duplicats de PCR

Un duplicat de PCR est défini par des coordonnées strictement identiques pour les deux *reads* de la paire et un score CIGAR identique. Dans le cas où la technique de préparation de la librairie génère des fragments aux extrémités aléatoires (donc hors amplicons), il faut éliminer les duplicats de PCR et ne tenir compte que d'un seul *read* (impact sur la

détection des mutations). Des outils tels que Picard, Samtools ou GATK peuvent réaliser cette étape.

2.3.2.2.4. Réalignement autour des Indels

Cette étape consiste à détecter les sites qui nécessitent un réalignement local (*indel* connues dans les bases de données, *indel* vues dans les alignements originaux et sites suggérant une *indel* cachée). Il y a alors recherche du meilleur alignement pour les sites ciblés en prenant en compte les différentes lectures s'alignant à ces positions, ce qui permet d'éliminer le maximum de faux positifs (Figure 21). Le réalignement peut être réalisé par GATK.

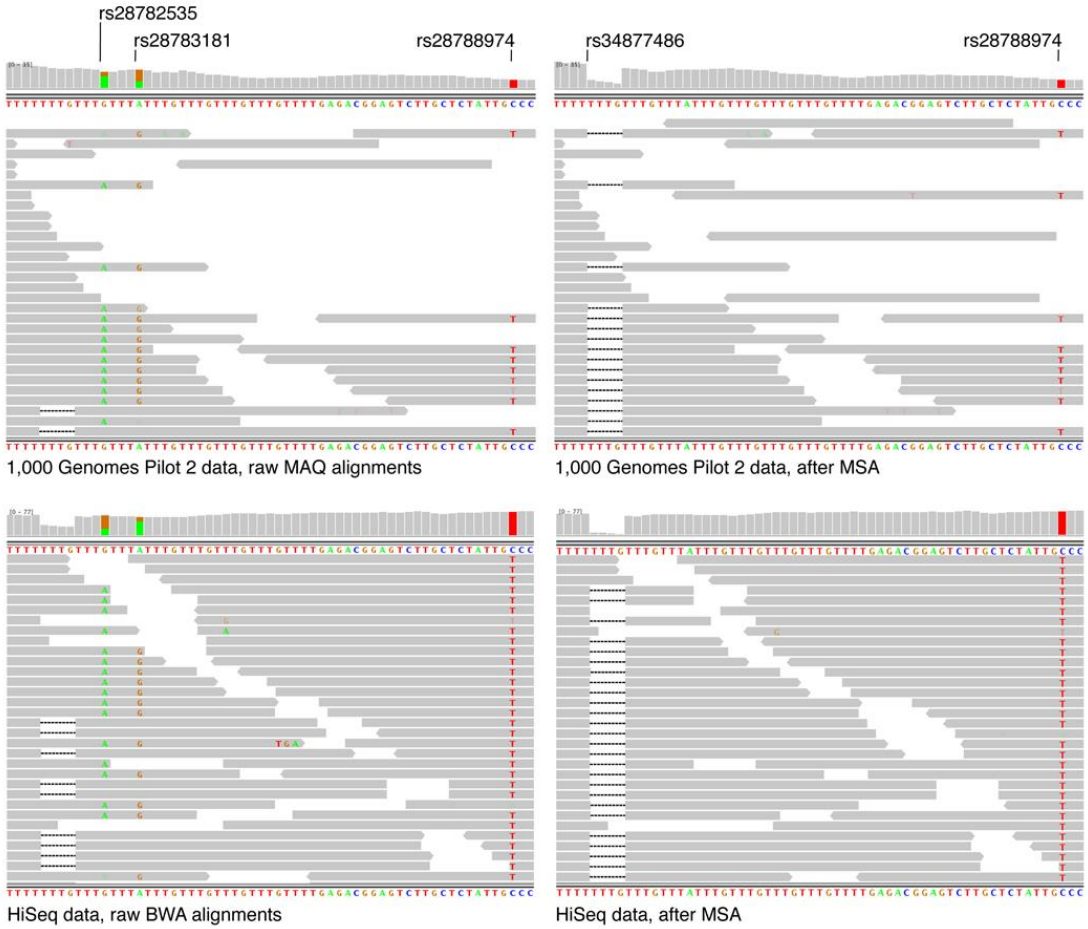

Figure 21. Effet du réalignement de séquence. A gauche : avant le réalignement et à droite : après le réalignement, suppression de variant correspondant à des faux positifs. MSA : *Multiple Sequence Alignement.* (31).

2.3.2.2.5. *Variant Calling*

Le but est d'identifier et de lister les variants détectés. Ceci peut être réalisée par des outils comme GATK, Samtools ou Varscan. Le format de fichier obtenu est le VCF (*Variant Call Format*). Celui-ci possède au minimum huit champs fixés : le chromosome, la position du variant, l'identifiant (numéro Rs, identifiant COSMIC …), la base de référence, la base alternative, la qualité moyenne des bases alternatives et les critères de passage du filtre prédéfinie (Figure 22). Des champs optionnels permettent d'ajouter des données (32). Ce fichier permet ensuite de trier et filtrer les variants sur des critères choisis. Les résultats sont visualisés avec un tableur de type excel, ce qui permet une exploitation simple et intuitive des données.

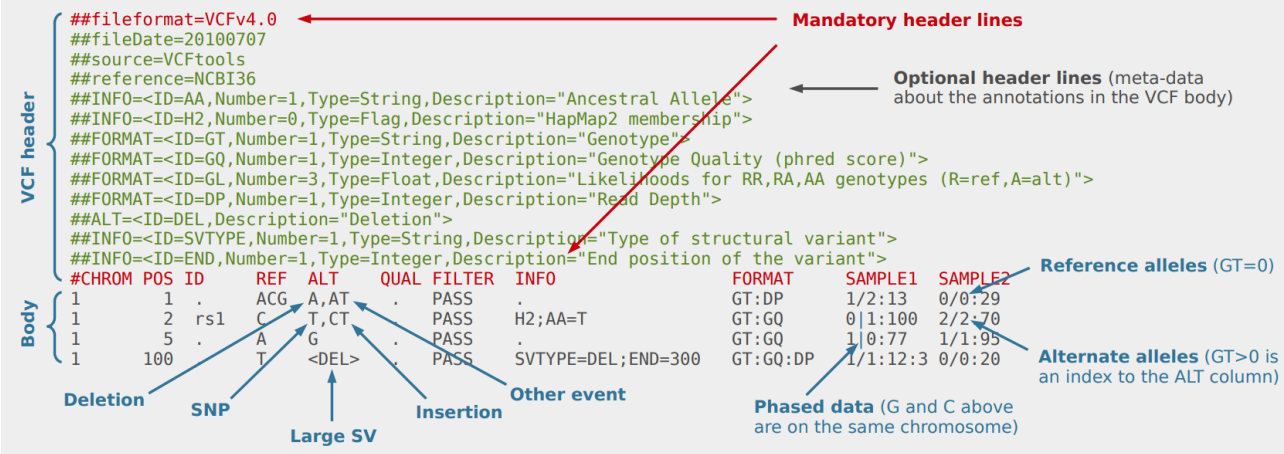

Figure 22. Exemple de fichier VCF et description des différents champs (32).

2.3.2.3. Analyse tertiaire

L'analyse tertiaire correspond à l'annotation des variants et à l'interprétation biologique des résutats et est détaillée dans le paragraphe 3.

3. Analyse tertiaire du NGS

3.1. Annotation des variants

L'annotation du fichier VCF est réalisé de façon systématique en fin de processus d'analyse des données et permettra de sélectionner le variant d'intérêt. L'objectif de l'annotation est de pouvoir ensuite filtrer les variants pour répondre à une question biologique donnée. Les variants du fichier VCF sont annotés grâce à l'interrogation de différentes bases de données ou encore des résultats des outils de prédiction. Plusieurs outils sont capables de réaliser cette annotation, tel que *Seattle Seq Annotation,* SnpEff,

Annovar ou encore VEP (*Variant Effect Predictor)* qui permet de réaliser des prédictions *in silico* sur plusieurs outils simultanément.

3.2. Classement des variants

L'ACMG rappelle dans ses recommandations de 2015 de ne plus utiliser les termes « mutation » ou « polymorphisme » mais seulement le terme « variant » qui sera interprété selon le niveau de pathogénicité (33). Un variant est donc une modification par rapport à la séquence de référence et les variants sont répartis en cinq classes différentes, allant de « pathogène » à « bénin ».

SNP Type	No. of SNPs
Nongene	2,255,102
Gene	1,165,204
Intron	1,064,655
Promoter	60,075
3′ UTR	16,350
5′ UTR	3,517
Splice regulatory site	2,089
Splice site	112
Synonymous	9,337
Stop→stop	17
Nonsynonymous	9,069
Stop→gain	121
Stop→loss	27
Total	3,420,306

Figure 23. Exemple du nombre de variants retrouvés pour un séquençage entier du génome humain (34).

A l'échelle d'un génome entier, on retrouve plus de trois millions de variations (Figure 23). Cependant, la grande majorité de ces variations n'auront pas d'effet délétère. Il faudra donc identifier parmi l'ensemble de ces variations celle(s) associée(s) au phénotype du patient. L'ACMG propose une méthode de classification des variants (Figure 24) en pondérant les principaux critères utilisés pour aboutir à la classification en cinq niveaux. Un résumé des critères de classification est fourni en annexe.

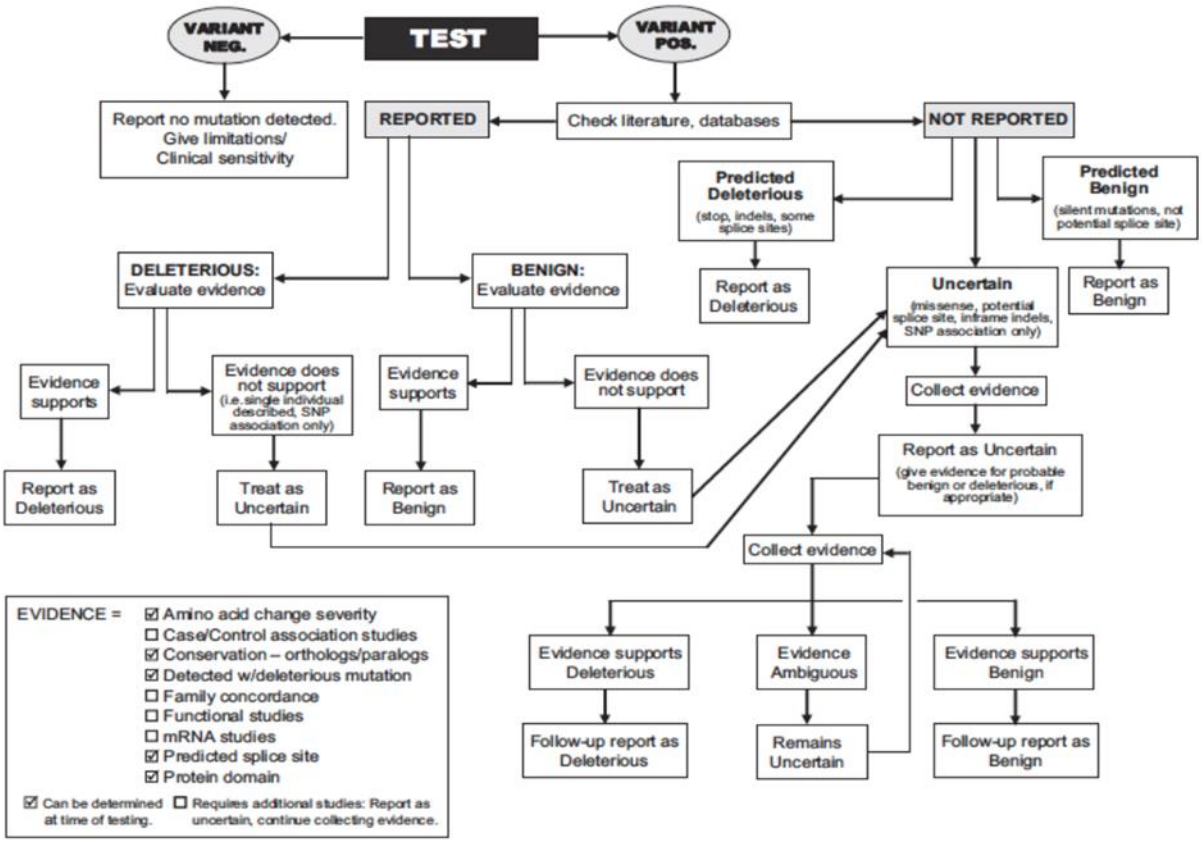

Figure 24. Organigramme décisionnel sur l'interprétation des variants identifiés en génétique constitutionnelle proposé par l'*American College Of Medical Genetics* (33).

3.2.1. Variants bénins et probablement bénins (classes 1 et 2)

Il s'agit de variants qui entrainent peu ou pas de modifications sur la protéine codée. Ces variants n'ont que des conséquences minimes voir aucune conséquence sur la protéine en aval. Il peut s'agir de polymorphisme (fréquence > 1% dans la population générale).

3.2.2. Variants de signification indéterminée (VSI) (classe 3)

Ces variants ne peuvent pas être interprétés dans l'état actuel des connaissances. La littérature, les bases de données et les outils de prédiction n'ont pas permis de les classer ou les prédictions sont contradictoires. Il n'est pas possible de les relier à la pathologie ni d'exclure leur implication.

3.2.3. Variants pathogènes et probablement pathogènes (classes 4 et 5)

Ces variants peuvent concerner des mutations non-sens entrainant l'apparition d'un codon stop prématuré, les insertions ou délétions qui entrainent un décalage du cadre de lecture et les remaniements de grande taille ainsi que les variants décrits pathogènes dans la

littérature et les bases de données. Ces variants entrainent des conséquences sur la protéine et peuvent être responsables de la pathologie et du phénotype.

3.3. Sélection des variants

3.3.1. Critères analytiques

La sélection des variants d'intérêt se fait tout d'abord sur des critères analytiques, afin d'éviter les faux positifs. La visualisation du fichier BAM peut permettre d'éviter des erreurs, tels que des variants retrouvés uniquement sur un seul sens de lecture, des éventuelles amorces non trimmées ou des faux positifs liés au pipeline d'analyse. Le fait de référencer l'ensemble des résultats dans une base de données locale permet de mettre en évidence des variants récurrents correspondant à des artéfacts. Une faible fréquence allélique peut correspondre à un artéfact ou une mosaïque (35). Des filtres sur la profondeur peuvent permettre d'exclure les variants avec de faibles profondeurs qui pourraient correspondre à des artefacts, le taux d'erreur du SHD est de l'ordre de 0,1 à 1%. Une vérification par une technique de référence type séquençage Sanger peut être réalisée.

3.3.2. Fréquences, polymorphisme et bases de données

La première étape est de déterminer si ce variant est référencé dans les bases de données ou décrit dans la littérature.

3.3.2.1. Bases de données

L'avènement du SHD et son accessibilité a permis le séquençage de plus en plus d'individus au fur et à mesure des années. Ceci a permis de constituer des bases de données référençant de plus en plus de variants et permettant une meilleure caractérisation des variants retrouvés en SHD.

3.3.2.1.1. Bases de données générales

Plusieurs projets de séquençage à grande échelle tel que le projet «1000 Genomes» ont été mené et ont permis la constitution des bases de données. La base ESP (*Exome Sequencing Project*) comprend actuellement les exomes de 6502 individus. La base ExAC (*Exome Aggregation Consortium*) regroupe les exomes de 60 706 individus (36) soit 14 cohortes de patients et 6 groupes de populations (Tableau 2). La base gnomAD (*Genome Aggregation Database*) correspond à une évolution d'ExAC et compile actuellement 123136 exomes et 15946 génomes. Elle regroupe les données d'ExAC et d'autres projets de séquençage tels que 1000 Genomes et HapMap. La base dbSNP est à,l'origine une

base de polymorphismes mais aujourd'hui elle référence également les variations selon leur degré de pathogénicité. Ces différentes bases de données vont permettre de filtrer les variants selon leur fréquence dans la population générale. On considère qu'un variant est un polymorphisme lorsque sa fréquence dans la population générale est supérieure à 1%. Cependant les informations issues de ces bases de données doivent être utilisées avec prudence. Ces bases ne contiennent pas que des variants bénins et les cohortes peuvent contenir des individus présentant diverses pathologies de l'adulte, seuls les phénotypes pédiatriques sévères étant éliminés. Les filtres utilisés avec ces bases de données ne sont donc pas parfaits et peuvent éliminer des variants pathogènes qu'ils considèrent comme bénins (faux négatifs).

Consortium/Cohort	Samples
1000 Genomes	1,851
Bulgarian Trios	461
GoT2D	2,502
Inflammatory Bowel Disease	1,675
Myocardial Infarction Genetics Consortium	14,622
NHLBI-GO Exome Sequencing Project (ESP)	3,936
National Institute of Mental Health (NIMH) Controls	364
SIGMA-T2D	3,845
Sequencing in Suomi (SISu)	948
Swedish Schizophrenia & Bipolar Studies	12,119
T2D-GENES	8,980
Schizophrenia Trios from Taiwan	1,505
The Cancer Genome Atlas (TCGA)	7,601
Tourette Syndrome Association International Consortium for Genomics (TSAICG)	297
Total	60,706

Population	Male Samples	Female Samples	Total
African/African American (AFR)	1,888	3,315	5,203
Latino (AMR)	2,254	3,535	5,789
East Asian (EAS)	2,016	2,311	4,327
Finnish (FIN)	2,084	1,223	3,307
Non-Finnish European (NFE)	18,740	14,630	33,370
South Asian (SAS)	6,387	1,869	8,256
Other (OTH)	275	179	454
Total	33,644	27,062	**60,706**

Tableau 2. Exomes constituants la base de données ExAC.

3.3.2.1.2. Bases de variants associés à un phénotype

Il existe des bases de données spécifiques de pathologies et issues de réseaux nationaux. C'est par exemple le cas de la base de données françaises des gènes du syndrome de Lynch (37). L'avantage de telles bases de données est leur fiabilité grâce à la validation de variants par des groupes d'expert, cependant elles comportent des cohortes plus restreintes que les bases internationales et sont donc moins exhaustives. La base ClinVar compile des soumissions de variants réalisées directement par les laboratoires ou par des groupes d'experts (38).

3.3.2.1.3. Publications

La revue de la littérature peut aider à caractériser des variants. Les sources sont documentées mais la recherche de telles informations se révèle longue et fastidieuse. Il existe cependant des bases référençant les différentes publications. La base HGMD (*Human Gene Mutation Database*) résulte de l'extraction et de la compilation des données de la littérature. C'est une base relativement complète. OMIM (Online Mendelian Inheritance in Man) recense plus de 3800 maladies secondaires à des mutations dans plus de 3400 gènes. Les données d'OMIM sont issues de la littérature et mises à jour régulièrement. Cependant cette base est loin d'être exhaustive due à la complexité de la collecte des informations (39).

Base de données	**Site Web**
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
Ensembl	http://www.ensembl.org/index.html
HapMap	http://www.hapmap.org
Human Gene Mutation Database	http://www.hgmd.cf.ac.uk/ac/index.php
OMIM	http://www.ncbi.nlm.nih.gov/omim/
UMD	http://www.umd.be
gnomAD	http://http://gnomad.broadinstitute.org/

Tableau 3. Principales bases de données des mutations de maladies génétiques.

3.3.3. Prédictions *in silico*

Lorsque le variant est absent des différentes bases de données ou que son implication dans la pathologie n'est pas clairement définie, des outils supplémentaires peuvent être utilisées. L'effet de la mutation sur l'épissage ou sur la protéine en aval peut alors être prédit par des algorithmes bioinformatiques. Les deux principales catégories de ces outils comprennent celles qui vont prédire si un changement faux-sens est préjudiciable à la fonction ou à la structure de la protéine résultante, et ceux qui prédisent les effets sur l'épissage.

3.3.3.1. Outils de prédiction

3.3.3.1.1. Prédiction de l'effet des variations faux sens

Une variation faux-sens est une altération ponctuelle d'un nucléotide qui aboutit à un changement d'acide aminé lors de la traduction. Le changement d'acide aminé peut avoir ou non des conséquences sur la fonction de la protéine produite et peut avoir des effets délétères ou non. En effet l'acide aminé peut être impliqué dans le métabolisme, dans une liaison avec un partenaire (ligand, ion métallique …), la communication intra ou inter cellulaire ou une implication dans les modifications post-traductionnelles. L'objectif est de statuer sur la modification de fonction de la protéine traduite.

3.3.3.1.1.1. Principes de fonctionnement

####### 3.3.3.1.1.1.1. Conservation de l'acide aminé

Un acide aminé conservé au cours de l'évolution a de grandes chances d'appartenir à un élément fonctionnel jouant un rôle important dans la protéine. L'alignement multiple est utilisé afin de visualiser la conservation des résidus entre les espèces (40). D'autre part, les acides aminés possèdent des structures physico-chimiques variés, influant sur la polarité et l'encombrement stérique. La conséquence prédite d'une mutation impliquant des acides aminés très différents d'un point de vue physico-chimique sera alors plus délétère que si elle concernait des acides aminés proches.

Par exemple, l'algorithme SIFT (*Sorting Intolerant From Tolerant*) part du postulat que des acides aminés importants seront conservés au cours de l'évolution, de sorte que les changements dans des positions bien conservées ont tendance à être prédits comme délétères. Par exemple, si une position donnée dans les alignements correspond toujours à l'isoleucine, on présume que la substitution par un autre acide aminé est aberrante et que l'isoleucine est nécessaire pour la fonction protéique. Par conséquent, une modification de tout autre acide aminé sera prédite comme délétère pour la fonction protéique. Si une position dans un alignement contient des acides aminés hydrophobes type isoleucine, valine et leucine, alors SIFT suppose, en effet, que cette position ne peut contenir que des acides aminés à caractère hydrophobe. Dans ce cas, les modifications correspondant à d'autres acides aminés hydrophobes sont généralement prédites comme étant tolérées, alors que les modifications correspondant à d'autres résidus (tels que chargés ou polaires) seront prédites comme affectant la fonction protéique, c'est-à-dire non tolérée.

Pour prédire si la substitution d'un acide aminé affecte la fonction protéique, SIFT considère la position à laquelle la variation s'est produite et le type d'acide aminé impliqué. Compte tenu d'une séquence protéique, SIFT choisit des protéines proches et obtient un alignement de ces protéines avec la protéine étudiée. Sur la base des acides aminés apparaissant à chaque position dans l'alignement, SIFT calcule la probabilité qu'un acide aminé soit toléré. Si cette valeur normalisée est inférieure à un seuil, la substitution est prédite non tolérée (41).

3.3.3.1.1.1.1.1. La matrice Grantham

Cette matrice est une matrice de transition entre les 20 acides aminés connus en se basant sur les propriétés physico-chimiques et le volume moléculaire de chaque acide aminé. Plus le score est petit plus les propriétés physico-chimiques de la protéine sont respectées. Le score de Grantham tente de prédire la différence entre deux acides aminés. Les scores de distance publiés par Grantham vont de 5 à 215. Un score faible reflète une faible distance évolutive. Un score élevé reflète une plus grande distance évolutive. Les scores élevés de Grantham sont considérés comme plus délétères. Ainsi, un remplacement de l'isoleucine pour la leucine, ou de la leucine pour l'isoleucine, a un score de 5 (et est considéré toléré). Une substitution de la cystéine par le tryptophane, ou du tryptophane par la cystéine, a un score de 215 et est prédit délétère (42).

3.3.3.1.1.1.1.2. Autres algorithmes de conservation

D’autres outils de prédiction se basent sur la conservation de blocs nucléotidiques, par utilisation d’une matrice d’intolérance aux changements de résidus c’est le cas de PhastCons, PhyloP, PROVEAN (PROtein Variation Effet Unalyzer) et GERP (*Genomic Evolutionary Rate Profiling)*.

3.3.3.1.1.1.2. Variation de structure

Certains outils de prédiction prennent en compte la structure de la protéine. Un variant peut avoir un effet délétère sur la structure si celui-ci touche des résidus critiques de la protéine. Parmi ces structures on peut citer par exemple les cœurs hydrophobes, les ponts disulfures ou encore les domaines fonctionnels connus.

Polyphen (*Polymorphism Phenotyping*) utilise à la fois des données d’alignements multiples, des annotations connues de séquences et des informations structurales provenant de banque de données de structure (PDB). Il donne la possibilité de prédire si la substitution engendre une déformation probable de la structure de la protéine par

utilisation d'un algorithme combinant HBPLUS (qui permet la prédiction de ponts hydrogènes entre chaînes latérales et entre chaines latérale et chaine principale) et Dist (qui calcule les contacts spatiaux d'un résidu) (43).

3.3.3.1.1.1.2.1. Autres outils

Les prédictions des conséquences des variants sur la structure de la protéine peuvent être également réalisées avec KD4v qui réalise des comparaisons de structure (44) ou encore avec la matrice de similarité BLOSUM62 (*BLOck Substitution Matrix*) qui permet de réaliser des alignements de séquences biologiques reliées évolutivement et de donner un score de similarité ou de ressemblance entre deux acides aminés (45).

3.3.3.1.1.1.3. Outils intégratifs

Des outils proposent de combiner plusieurs algorithmes de prédiction en permettant l'établissement d'un score global de prédiction. C'est par exemple le cas de Condel (*CONsensus DELereriousness score of missense SNVs*) qui intègre la sortie d'outils de calcul visant à évaluer l'impact des variations faux sens non synonymes sur la fonction des protéines et calcule une moyenne pondérée des scores de ces outils. Il intègre les résultats de cinq outils : SIFT, Polyphen2, MAPP, LogR-Pfam Valeur-E et MutationAssessor. D'autres outils utilisent le même principe de combinaison tel que MutationTaster qui combine les résultats de PhasCons, PhyloP, NNsplice et du score de Grantham. CADD (*Combined Annotation Dependant Depletion)* combine de multiples données sur la conservation et la tolérance de changement de résidus pour calculer le C-score (46–48).

3.3.3.1.1.2. Utilisation

Ces outils ne sont en général pas utilisés individuellement mais en parallèle. Il convient donc d'utiliser la combinaison d'outil la plus pertinente, et de ne pas utiliser des outils se basant sur les mêmes algorithmes de prédiction. Les entretiens avec les biologistes du CHRU de Lille montrent que l'utilisation en routine consiste à utiliser une combinaison de plusieurs logiciels et à vérifier si leurs prédictions sont concordantes.

Outil de prédiction	Site Web
SIFT	http://sift.jvci.org
POLYPHEN	http://genetics.bwh.harvard.edu/pph2
CADD	http://cadd.gs.washington.edu
PROVEAN	http://provean.jcvi.org/index.php
Condel	http://bg.upf.edu/condel/home
MutationTaster	http://www.mutationtaster.org
BLOSUM62	http://www.ensembl.info/ecode/blosum62
PhastCons / PhyloP	http://compgen.bscb.cornell.edu/phast/

Tableau 4. Principaux outils bioinformatique de prédiction des variants faux sens.

3.3.3.1.2. Prédiction des effets sur les transcrits (épissage)

L'épissage est la suppression d'introns réalisée par un complexe appelé spliceosome. Le spliceosome reconnaît le site d'épissage à travers un appariement de bases.

Les variants ayant des conséquences sur l'épissage se situent principalement sur les sites canoniques d'épissage. Les logiciels de prédiction calculent un score qui renseigne sur la force du site d'épissage et il y a ensuite comparaison des scores sauvages et de la séquence mutée pour prédire l'effet.

Le logiciel MaxENTScan se base sur une approche de modélisation de motifs de courtes séquences impliquées dans l'épissage de l'ARN qui prend simultanément en compte les dépendances non adjacentes et adjacentes entre les positions. Cette méthode est basée sur le «principe d'entropie maximum» et généralise la plupart des modèles probabilistes de motifs de séquences tels que les modèles de matrice de poids et les modèles de Markov inhomogènes (49). Le logiciel *Human Splicing Finder* se base sur des matrices de prédiction pour les sites de liaison des protéines d'épissage. Il combine 12 algorithmes différents pour identifier et prédire l'effet des mutations sur les motifs d'épissage, y compris les sites d'épissage donneur et accepteur, le point de branchement et les séquences auxiliaires connus pour renforcer ou réprimer l'épissage: *Exonic Splicing Enhancers* (ESE) et les *Exonic Splicing Silencers* (ESS) (50). Les performances de ces outils ont déjà été comparées précédemment dans la littérature afin de proposer des algorithmes d'utilisation

combinés (51). Ils se révèlent assez efficaces lorsque le variant est proche des sites consensus mais les performances s'amenuisent lorsque le variant s'éloigne de ces derniers.

Outil de prédiction	Site Web
Human Splicing Finder	http://www.umd.be/HSF
MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
NNSplice	https://omictools.com/nnsplice-tool

Tableau 5. Principaux outils bioinformatiques de prédiction des variants affectant l'épissage.

3.4. Confrontation aux données cliniques

3.4.1. Analyses de ségrégation

Les analyses de ségrégation ou de transmission familiale permettent de déterminer le caractère hérité ou non d'un variant. Dans l'idéal, il est nécessaire de pouvoir réaliser la recherche du variant chez les 2 parents du cas index (analyse de trio). Lorsque le variant n'est retrouvé chez aucun des deux parents sains, il est dit *de novo*, ce qui constitue un argument supplémentaire en faveur de l'implication dans la pathologie du patient atteint (cas index ou *propositus*). Lorsque le variant est hérité, Il faudra déterminer si d'autres patients de la famille sont atteints et si le variant ségrége avec la maladie dans la famille

3.4.1.1. Mode de transmission

L'analyse étendue de la famille permet de prédire le mode de transmission et d'orienter vers le ou les types de variants à rechercher (variant unique ou 2 variants sur un même gène, variant sur le chromosome X …).

3.4.1.1.1. Hérédité autosomique dominante

Les sujets atteints sont porteurs d'un allèle muté d'un gène situé sur un autosome. Les individus atteints peuvent être des hommes ou des femmes. Il n'y a pas de biais de transmission en fonction du sexe.

Le risque de transmission à la descendance pour un individu atteint est de 50%. Les variants pathologiques doivent être retrouvés chez d'autres membres atteints de la famille.

On recherchera donc les variants candidats dans la famille pour s'assurer de son absence chez les individus non atteints et sa présence chez les individus atteints (figure 25).

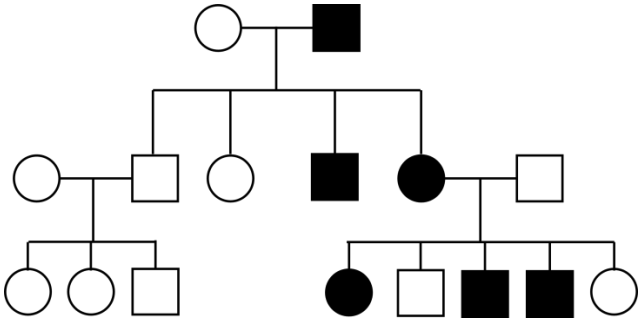

Figure 25. Exemple d'arbre généalogique d'une pathologie autosomique dominante (les carrés correspondent aux hommes, les ronds aux femmes, la couleur noire aux individus atteints).

3.4.1.1.2. Hérédité autosomique récessive

Les sujets atteints sont porteurs de 2 allèles mutés d'un gène localisé sur un autosome (homozygotes ou hétérozygotes composites). Les hétérozygotes sont non atteints (conducteurs). Il n'existe pas de différence de fréquence, ni de transmission en fonction du sexe. Les autres membres atteints de la famille seront généralement dans la fratrie du cas index et seront également porteurs de 2 mutations. Dans le cas des hétérozygotes composites, on recherchera les variants candidats chez les parents pour s'assurer qu'ils sont bien situés sur des allèles différents (figure 26).

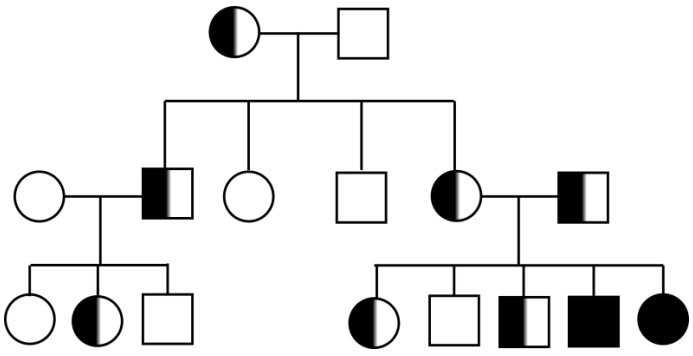

Figure 26. Exemple d'arbre généalogique d'une pathologie autosomique récessive (les carrés correspondent aux hommes, les ronds aux femmes, les formes totalement noires correspondant aux porteurs atteints (homozygotes ou hétérozygotes composites) et les formes en noire et blanc correspondent aux conducteurs non atteints (hétérozygotes).

3.4.1.1.3. Hérédité liée au chromosome X

Le gène incriminé se trouve sur le chromosome X (gonosome). Les sujets atteints sont hémizygotes (garçons) ou homozygotes (filles). Les filles hétérozygotes sont conductrices et non atteintes ou peuvent présenter un phénotype atténuée lié au biais de l'inactivation du X. Les femmes conductrices ont 50% des garçons atteints. Toutes les filles d'un homme atteint sont conductrices. Actuellement, on distingue plutôt de façon général l'hérédité liée à l'X sans préciser le caractère dominant ou non. (Figure 27).

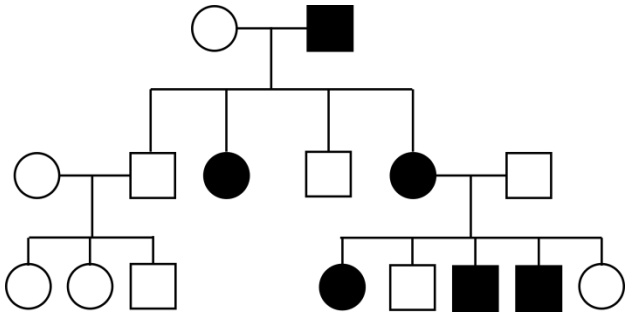

Figure 27. Exemple d'arbre généalogique d'une pathologie liée à l'X.

3.4.1.2. Mosaïcisme et pénétrance incomplète.

Un variant peut être présent à l'état de mosaïque chez un individus, c'est-à-dire présent seulement dans une partie des cellules. La pénétrance de la maladie doit être également prise en compte. La pénétrance est la probabilité pour un individu porteur du génotype à risque d'être atteint par la maladie. La pénétrance d'une maladie est complète (égale à 100%) quand tous les individus porteurs de l'allèle muté sont malades. On parle de pénétrance incomplète (< 100%) lorsque tous les porteurs du génotype à risque ne sont pas malades.

3.4.2. Réunions de Concertation Pluridisciplinaire

L'identification de la mutation en lien avec la pathologie nécessite la confrontation des données moléculaires avec le phénotype et donc une grande proximité entre le biologiste et son homologue clinicien, afin de sélectionner au mieux la mutation parmi tous les variants candidats. Pour cela, les réunions de concertation pluridisciplinaire, que ce soit au niveau local ou encore au niveau national sont très utiles pour confronter les données cliniques et biologiques. La constitution de réseaux de centres experts dans une pathologie donnée permet un partage d'expérience et des discussions autour des variants difficilement interprétables qui permettront parfois de les reclasser.

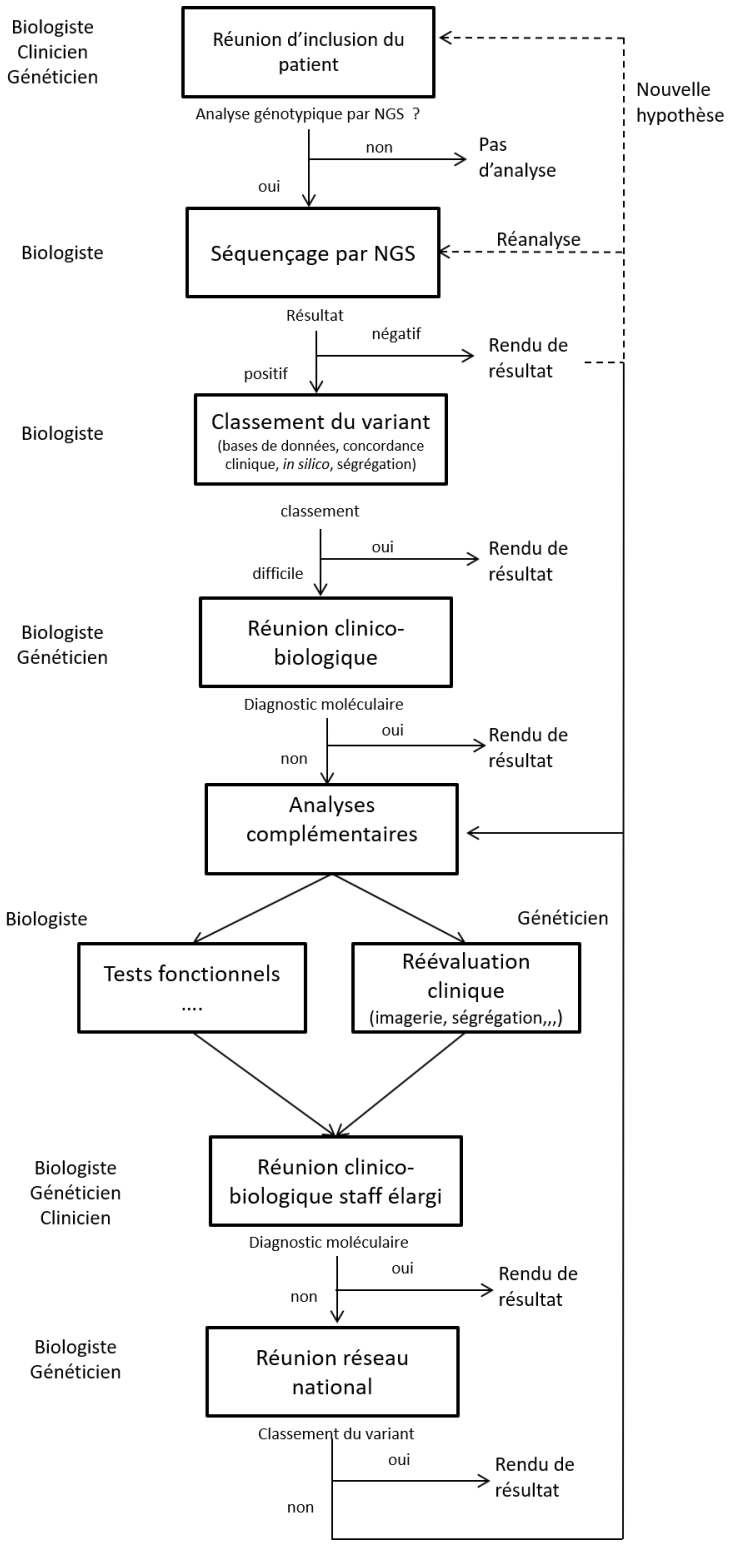

Figure 28. Logigramme de prise en charge pluridisciplinaire des analyses génétiques par NGS.

3.5. Conséquence de la classification des variants

La classification des variants peut avoir des conséquences majeures sur la prise en charge du patient et également des apparentés. En fonction de la problématique initiale, celles-ci peuvent prendre plusieurs formes. La découverte d'un variant pathogène peut entrainer une modification ou la mise en place d'une thérapeutique adaptée au trouble constitutionnel identifié, cette prise en charge pouvant être lourde, à la fois pour le patient mais également économiquement pour la société. Un conseil génétique peut être proposé c'est-à-dire l'information sur l'hérédité du trouble pour la descendance. En cas de variant pathogène identifié un diagnostic prénatal (DPN) ou un diagnostic pré-implantatoire (DPI) dans le cadre d'une procréation médicalement assistée (PMA) peut être proposé. Un test présymptomatique peut être également proposé aux apparentés lorsqu'un variant pathogène est identifié dans la famille. La classification d'un variant à tort peut donc entrainer des conséquences lourdes sur le patient et sa famille.

4. Objectifs du travail

L'objectif de ce travail est de réaliser une étude sur l'intérêt des outils d'aide à l'interprétation des variants identifiés en NGS en génétique constitutionnelle au CHRU de Lille. La première partie de ce travail consiste à déterminer la performance des outils *in silico* utilisés en pratique diagnostique. En effet, comme dit précédemment, plusieurs logiciels sont utilisés en parallèle puis le biologiste évalue la pertinence de la prédiction par concordance de celle-ci entre l'ensemble des logiciels utilisés. Cependant nous n'avons que très peu d'informations sur les performances de ces outils. L'objectif est donc de proposer le meilleur choix de logiciels en fonction de la question posé et de pouvoir avoir un avis critique sur les prédictions qui en ressortent. La deuxième partie de ce travail fait le point sur les usages et perspectives des tests fonctionnels au CHRU de Lille. Le but est de proposer des axes d'amélioration et de faire le point sur les besoins actuels et futurs au sein de notre plate-forme de biologie moléculaire.

II. Etude des performances des outils de prédiction

La rencontre avec de nombreux biologistes impliqués dans le diagnostic moléculaire par NGS au Centre de Biologie et Pathologie (CBP) du CHRU de Lille a permis de faire un point sur les pratiques de séquençage et d'interprétation des variants identifiés en NGS en génétique constitutionnelle. L'ensemble des biologistes interrogés utilisent des algorithmes

de prédiction en aide à l'interprétation des variants retrouvés en NGS. Ces outils de prédiction *in silico* ont une part plus ou moins importante dans les décisions d'interprétation. Cependant, selon les recommandations de l'ACMG, la cotation des outils de prédiction est faible sur la décision de classification finale (16). La pratique locale consiste à prédire l'effet du variant à l'aide de plusieurs logiciels et de s'assurer que les prédictions sont concordantes entre elles. Dans le cas de discordance, aucun logiciel n'est préféré à un autre. Nous souhaitons donc étudier la performance de quelques outils de prédiction à l'aide d'un panel de variants bien caractérisés retrouvés au laboratoire ou issus de la littérature. Nous avons réalisé cette étude sur les outils prédisant les effets des mutations faux sens sur les protéines. Le but est de proposer une stratégie de bonne pratique d'utilisation de ces outils.

1. Matériels et méthodes

1.1. Recueil des données

Pour réaliser notre étude de performance des outils de prédiction *in silico* de l'effet des variations faux sens, nous souhaitons tester ces outils sur des variants faux sens classés bénins ou pathogènes. Un recueil a donc été réalisé dans trois laboratoires afin d'avoir un nombre de variants suffisamment important pour réaliser une étude de performance.

1.1.1. Sets de variants

Trois sets de données, correspondant à 3 thématiques différentes ont été utilisés, un set de variants de gènes d'oncogénétique digestive (OD), un set de variants de gènes des rétinopathies (RP) et un set de variants de gènes de déficience intellectuelle (DI) (Tableau 6). Dans cette étude, nous nous limitons aux variants de classe 1 (bénin) et de classe 5 (pathogène).

Pathologie	Oncogénétique Digestive OD	Rétinites pigmentaires (RP)	Déficience intellectuelle (DI)
Gènes étudiés	*MLH1, MSH2, MSH6, MUTYH, APC*	*PRPH2, RHO, NR2E3, ABCA4, ELOVL4*	*MED13L, GRIN2B*
Nombre de variants pathogènes	31	28	70
Nombre de variants bénins	37	14	39

Tableau 6. Gènes et variants utilisés pour l'étude de performance

1.1.1.1. Oncogénétique digestive

Concernant l'oncogénétique digestive, 5 gènes impliqués dans le syndrome de Lynch ou la polypose adénomateuse et pour lesquels il existe une base de données nationale, ont été sélectionnés (*MLH1, MSH2, MSH6, MUTYH* et *APC)*. Les gènes impliqués ont été identifiés il y a de nombreuses années (entre 1993 et 1995 pour les gènes MMR, 1991 pour *APC* et 2002 pour *MUTYH*) Il existe donc un recul important pour ces gènes, des réseaux et groupes de travail nationaux et internationaux et des bases de données spécifiques issues de différentes sociétés savantes (*International Society for Gastrointestinal Hereditary Tumours InSight,* Réseau Français des Laboratoires d'Oncogénétique Digestive …). Les variants sont classés selon des décisions collégiales établies au sein de ces réseaux prenant en compte de nombreux critères dont le phénotype clinique des patients, la récurrence des variants et les données de la littérature. Les variants sélectionnés pour ce travail ont tous été identifiés au laboratoire et leur classement en pathogène ou bénin a été validé par le Réseau National des Laboratoires d'Oncogénétique Digestive. Les algorithmes de prédiction ont un poids faible dans la décision finale de classification des variants.

1.1.1.2. Rétinopathies

Les rétinopathies (RP) représentent à un groupe de pathologies dégénératives de la rétine qui se caractérisent par une perte progressive et graduelle de la vision évoluant généralement vers la cécité. Elles résultent d'anomalies génétiques des photorécepteurs, cônes ou batônnets. Il s'agit de maladies génétiques complexes : tous les modes de transmission ont été décrits. Elles sont génétiquement hétérogènes avec plus d'une cinquantaine de gènes identifiés comme responsables de ces pathologies et peuvent être associées à d'autres manifestations cliniques dans des syndromes différents et variés (Syndrome de Usher, Syndrome de Bardet Biedl …).

Pour la thématique RP, les variants sont issus de 5 gènes et ont été identifiés au laboratoire (*PRPH2, RHO, NR2E3, ABCA4* et *ELOVL4)*. Il existe une base de données des gènes de rétinopathie aidant à l'interprétation des variants (https://sph.uth.edu/retnet/). Lorsque le variant n'est pas répertorié dans cette base, d'autres bases sont utilisées (HGMD, Clinvar, ExAc). Les outils d'interprétation *in silico* inclus dans Alamut sont également utilisés en support au classement des variants. Les variants difficiles d'interprétation sont discutés en réunion clinico-biologiques locale ou via un centre de référence national et sont classés à la fois sur des critères clinico-biologiques forts, une présence pertinente dans la littérature ou dans les bases de données spécifiques de pathologies et le cas échéant des analyses de ségrégation. Les outils de prédiction *in silico* ont donc une importance moyenne dans la classification des variants dans cette thématique.

1.1.1.3. Déficience intellectuelle

Les déficiences intellectuelles non syndromiques sont très hétérogènes cliniquement et génétiquement ; le séquençage d'exomes de patients atteints a permis d'identifier plus de 500 gènes responsables de ces pathologies (52).

Concernant la DI, deux gènes ont été sélectionnés pour ce travail, *MED13L* et *GRIN2B*. L'implication de ces gènes dans la déficience intellectuelle est récente (après 2010) et la physiopathologie et la fonction des protéines codées est encore mal connue. Il y a encore assez peu de variants exploitables au laboratoire pour un même gène. En raison du manque de données dans la littérature et dans les différentes bases, la méthode de classification est très dépendante des méthodes de prédiction *in silico* qui ont un poids important dans la décision. Les variants pathogènes de *MED13L* ont été sélectionnés

dans la littérature, dans les bases de données et certains ont été identifiés au laboratoire, les variants pathogènes de *GRIN2B* sont issus de la littérature (53–55). Nous n'avons pas pu extraire une liste de variants bénins identifiés au laboratoire. Ceux-ci ont donc été extraits de la base de données GnomAD (http://gnomad.broadinstitute.org/) et ont été choisis selon leur fréquence (retrouvés plus de 150 fois).

1.2. Prédictions *in silico*

Les sets de variants extraits ont été convertis en VCF. Dans les comptes rendus diagnostiques, la nomenclature utilisée pour les variants correspond à leur position sur le cDNA. Il est donc nécessaire de convertir l'ensemble des positions cDNA en positions génomiques afin de les intégrer dans le fichier VCF. Pour réaliser ces transformations, j'ai utilisé à la fois l'outil de conversion de la base Mutalyzer (https://mutalyzer.nl/) et une macro Excel pour obtenir le fichier VCF correspondant. Le fichier obtenu est alors analysé avec l'outil VEP qui permet de lancer simultanément plusieurs outils de prédictions *in silico* (Variant Effect Predictor, www.ensembl.org/vep). Nous avons réalisé cette étude de comparaison sur les 9 logiciels de prédiction *in silico* les plus utilisés au CHRU (Tableau 7). Ces 9 logiciels ont été choisis après entretien avec les différents biologistes du Centre de Biologie et Pathologie de Lille.

Outils	Méthode de prédiction	Scoring obtenu via VEP (de bénin à pathogène)	Se attendue	Spe attendue
Condel	Combinatoire : SIFT, Polyphen2, MAPP, LogR-Pfam Valeur-E et MutationAssessor	0 à 1	NR	NR
CADD	Calcul du c-score : conservation physico-chimique et conservation inter-espèce	0 à 1	0.93	0.46
SIFT	Conservation inter-espèce	1 à 0	0.83	0.86
PolyPhen2	Conservation structurale de la protéine et conservation inter-espèce	0 à 1	0.86	0.82
GERP	Conservation nucléotidique multi-espèces	0 à 1	NR	NR
BLOSUM	Matrice d'intolérance, écart entre acides aminés	-3 à 3	NR	NR
Phast Cons	Conservation phylogénétique de blocs nucléotidiques	0 à 1	NR	NR
Mutation Taster	Combinatoire : Grantham, PhasCons, PhyloP, ExAc, NNsplice	0 à 1	0.88	0.87
PROVEAN	Calcul d'un score d'alignement des changements de résidus et conservation	0 à 1	0.78	0.78

Tableau 7. Outils de prédiction *in silico* utilisés au CHRU de Lille, méthode de prédiction, échelle de scores obtenus avec VEP et performances annoncées par les développeurs. NR pour données Non Renseignées. Se : Sensibilité, Spe : Spécificité.

1.3. Analyses statistiques

Les analyses statistiques ont été réalisées à l'aide du logiciel R v3.4.3 (www.r-project.org). Les graphes obtenus ont été réalisés grâce au package « ggplot2 ». Le script complet rédigé est fourni en annexe. Deux méthodologies permettant la comparaison des performances des logiciels ont été utilisées.

Nous avons évalué les logiciels individuellement par la détermination des paramètres de performance (sensibilité, spécificité, valeur prédictive positive, valeur prédictive négative). Etant donné que chaque logiciel fournit en sortie d'analyse, un score propre à lui-même et qu'il n'existe pas de seuil de pathogénicité prédéterminé pour ce score, il est nécessaire de déterminer pour chaque logiciel le seuil idéal permettant la meilleure discrimination possible entre variant pathogène et bénin. Pour cela, les sets de données sont utilisés pour tracer les courbes ROC pour chaque logiciel (*Receiver Operating Characteristic*) (56). Elle est obtenue par le calcul de la sensibilité et de la spécificité pour chaque seuil du logiciel et la représentation graphique est obtenue sous la forme d'une courbe qui donne la sensibilité en fonction du taux de faux positifs (1-Spécificité). Le calcul de l'AUC (aire sous la courbe) de chaque courbe ROC est réalisé afin de s'assurer que sa valeur est supérieure à 0,5, une valeur d'AUC inférieure à 0,5 correspondant à des prédictions liées au hasard. Le seuil idéal correspond au point le plus proche de l'idéal (1;1) ou au point le plus loin de la diagonale (Figure 29).

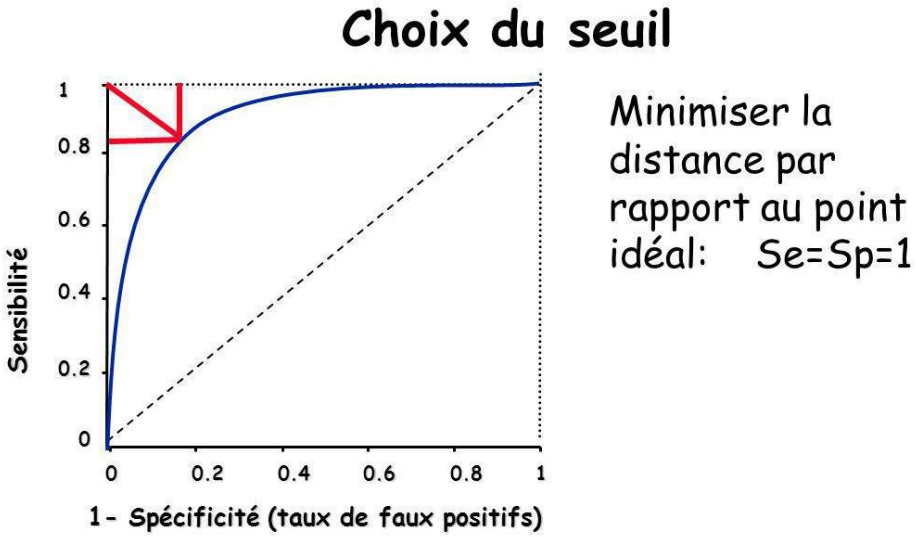

Figure 29. Méthode de détermination du seuil idéal grâce à la courbe ROC.

A partir du seuil idéal, nous avons pu déterminer pour chaque logiciel, les valeurs des paramètres de performance. La sensibilité est la probabilité d'obtenir un score qui prédit un effet délétère pour un variant pathogène. La spécificité est la probabilité d'obtenir un score qui prédit un effet non délétère pour un variant bénin (57). La valeur prédictive positive est la probabilité que le variant soit pathogène lorsque le logiciel prédit un effet délétère. La valeur prédictive négative est la probabilité que le variant soit bénin lorsque le logiciel prédit un effet non délétère. Les formules utilisées sont présentées dans le tableau 8. Les résultats ont été compilés sur le logiciel R afin d'être exploité graphiquement.

Score de Performance	Sensibilité	Spécificité	Valeur Prédictive Positive	Valeur Prédictive Négative
Formule	$\frac{VP}{VP+FN}$	$\frac{VN}{VN+FP}$	$\frac{VP}{VP+FP}$	$\frac{VN}{VN+FN}$
Formule de l'intervalle de confiance à 95%.	$Se \pm 1.96\sqrt{\frac{Se(1-Se)}{n1}}$	$Spe \pm 1.96\sqrt{\frac{Spe(1-Spe)}{n2}}$	$VPP \pm 1.96\sqrt{\frac{VPP(1-VPP)}{n1}}$	$VPN \pm 1.96\sqrt{\frac{VPN(1-VPN)}{n2}}$

Tableau 8. Formules utilisées pour calculer les scores de performance. VP : Vrais Positifs ; VN : Vrais Négatifs ; FP : Faux positifs ; FN : Faux Négatifs. VPN, Se : Sensibilité, Spe : Spécificité, n1 : nombre de variants pathogènes ; n2 : nombres de variants bénins ; Valeur Prédictive Négative, VPP : Valeur Prédictive Positive.

La méthode CART (*Classification And Regression Trees*) est également utilisée. Il s'agit d'une méthode de classification par arbre (58). Elle permet une évaluation cette fois de l'utilisation des logiciels entre eux et non pas individuellement comme la méthode ROC. Cette méthode permet de déterminer la meilleure association de logiciels discriminant correctement le maximum de variants.

La construction d'un arbre de discrimination ou régression consiste à déterminer une séquence de nœuds. Un nœud est défini par le choix conjoint d'une variable parmi les explicatives et d'une division qui induit une partition en deux classes. Implicitement, à chaque nœud correspond donc un sous-ensemble de l'échantillon auquel est appliquée une dichotomie. Une division est elle-même définie par une valeur seuil de la variable quantitative sélectionnée ou un partage en deux groupes des modalités si la variable est qualitative.

À la racine ou nœud initial correspond l'ensemble de l'échantillon ; la procédure est ensuite itérée sur chacun des sous-ensembles, à chaque nœud ne sont gardés que les sous-ensembles pas encore classés. Ici les variables sont les scores de décisions des logiciels de prédiction avec leurs seuils et la modalité binaire est le caractère bénin ou pathogène des variants (Figure 30).

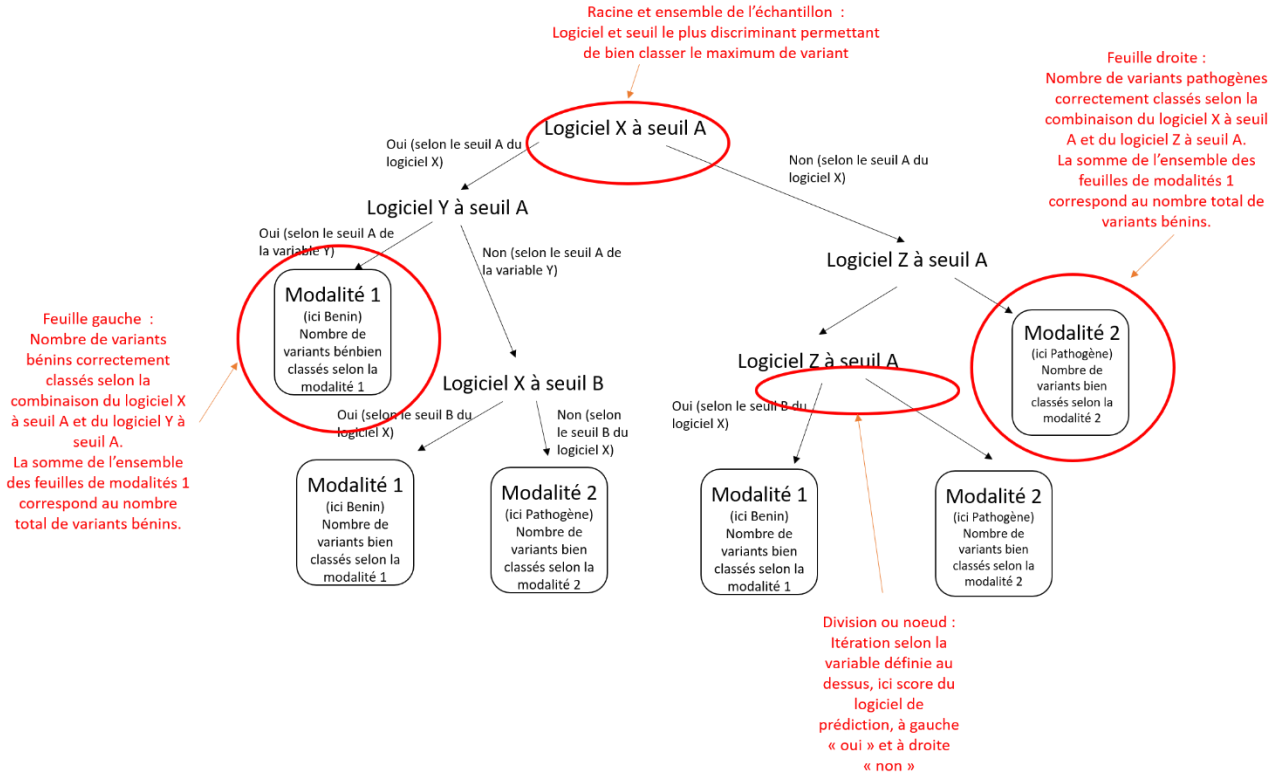

Figure 30. Mode de représentation d'un arbre selon la méthode CART et notice explicative des différents éléments obtenus.

2. Résultats

2.1. Variants issus du set oncogénétique digestive

2.1.1. Performances individuelles des logiciels

Les courbes ROC obtenues pour les 9 outils montrent des performances qui ne sont pas dues au hasard (AUC>0,5) (Figure 31). Ces courbes ROC ont permis de déterminer pour chaque logiciel les seuils optimaux permettant d'obtenir les meilleures performances. Les valeurs de performances ainsi que leur intervalle de confiance à 95% ont été rapportées dans le tableau 9. Les 2 logiciels présentant les meilleures performances individuelles globales (Se,Spe, VPP, VNP et AUC) sont Condel et CADD.

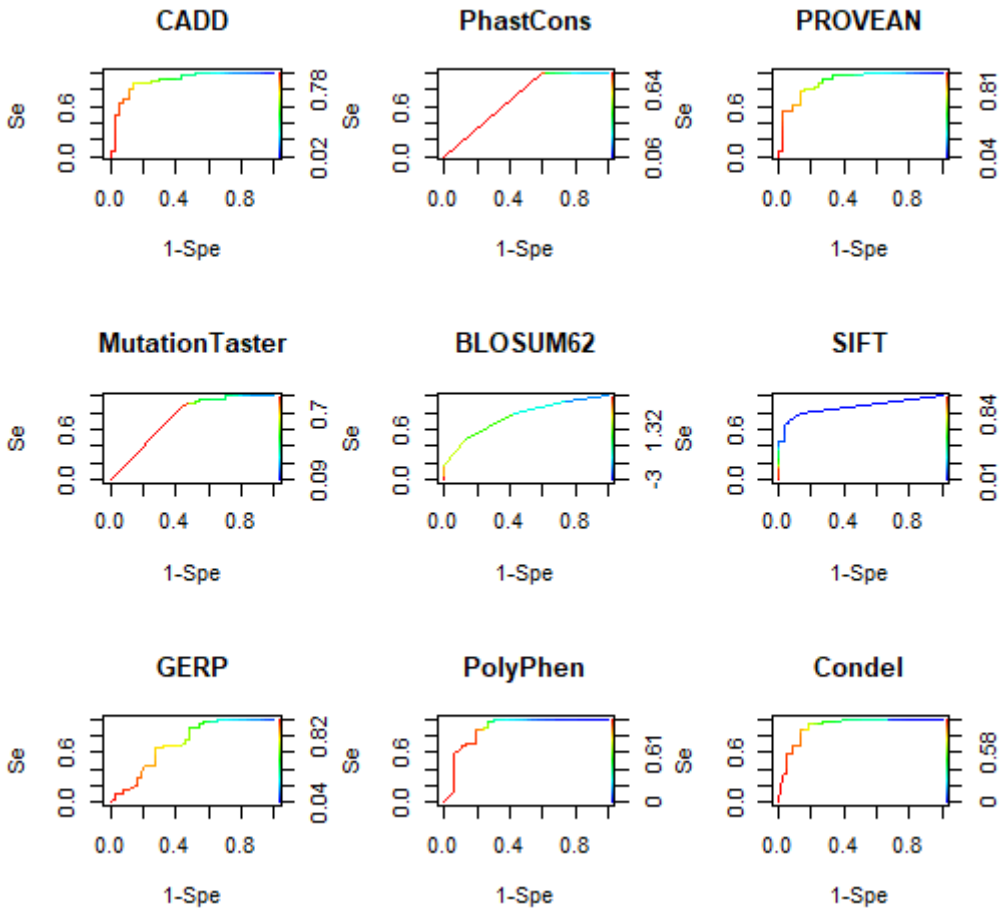

Figure 31. Courbes ROC obtenues pour les 9 outils de prédiction sur les variants du set oncogénétique digestive.

Logiciel	AUC	Se	Spe	VPP	VPN
Condel	0,91	0,84 (0,71-0,97)	0,86 (0,75-0,97)	0,83 (0,71-0,97)	0,86 (0,75-0,97)
SIFT	0,86	0,94 (0,85-1)	0,70 (0,56-0,84)	0,73 (0,57-0,88)	0,93 (0,84-1)
PolyPhen	0,89	0,84 (0,71-0,97)	0,81 (0,68-0,93)	0,79 (0,64-0,93)	0,86 (0,74-0,97)
GERP	0,71	0,61 (0,44-0,78)	0,73 (0,59-0,87)	0,66 (0,49-0,82)	0,69 (0,54-0,84)
BLOSUM62	0,76	0,87 (0,75-0,99)	0,49 (0,33-0,65)	0,59 (0,41-0,76)	0,82 (0,69-0,94)
PROVEAN	0,89	0,77 (0,63-0,92)	0,84 (0,72-0,96)	0,80 (0,66-0,94)	0,82 (0,69-0,94)
MutationTaster	0,74	1,00 (1-1)	0,24 (0,10-0,38)	0,53 (0,35-0,70)	1,00 (1-1)
PhastCons	0,70	1,00 (1-1)	0,41 (0,25-0,56)	0,58 (0,41-0,76)	1,00 (1-1)
CADD	0,90	0,84 (0,71-0,97)	0,86 (0,75-0,97)	0,84 (0,71-0,97)	0,86 (0,74-0,97)

Tableau 9. AUC (aire sous la courbe) des courbes ROC, performances des logiciels de prédiction et intervalles de confiance à 95% déterminés à partir des variants du set oncogénétique digestive. Se : Sensibilité, Spe : Spécificité, VPN : Valeur Prédictive Négative, VPP : Valeur Prédictive Positive.

2.1.2. Performances croisées

L'arbre de classement obtenu (méthode CART) permet de reclasser l'ensemble des variants grâce à l'utilisation conjointe des logiciels Condel, Polyphen, CADD et GERP. Condel est le logiciel le plus discriminant, en second niveau arrivent Polyphen et GERP (Figure 32). La combinaison de Condel, Polyphen et GERP semble donc être celle qui présente le plus d'intérêt dans la thématique OD.

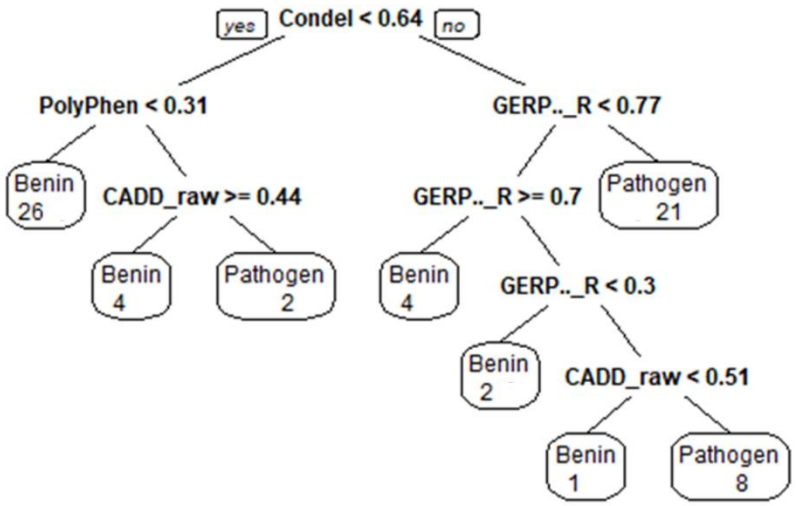

Figure 32. Arbre de classement obtenu par la méthode CART sur les variants du set oncogénétique digestive.

2.2. Variants issus du set rétinopathies

2.2.1. Performances individuelles des logiciels

Les courbes ROC obtenues pour les 9 outils montrent des performances qui ne sont pas dues au hasard (AUC>0,5) (Figure 33). Les valeurs de performances obtenues ainsi que leur intervalle de confiance à 95% sont présentées dans le tableau 10. Les 2 logiciels présentant les meilleures performances individuelles globales sont Condel et PROVEAN.

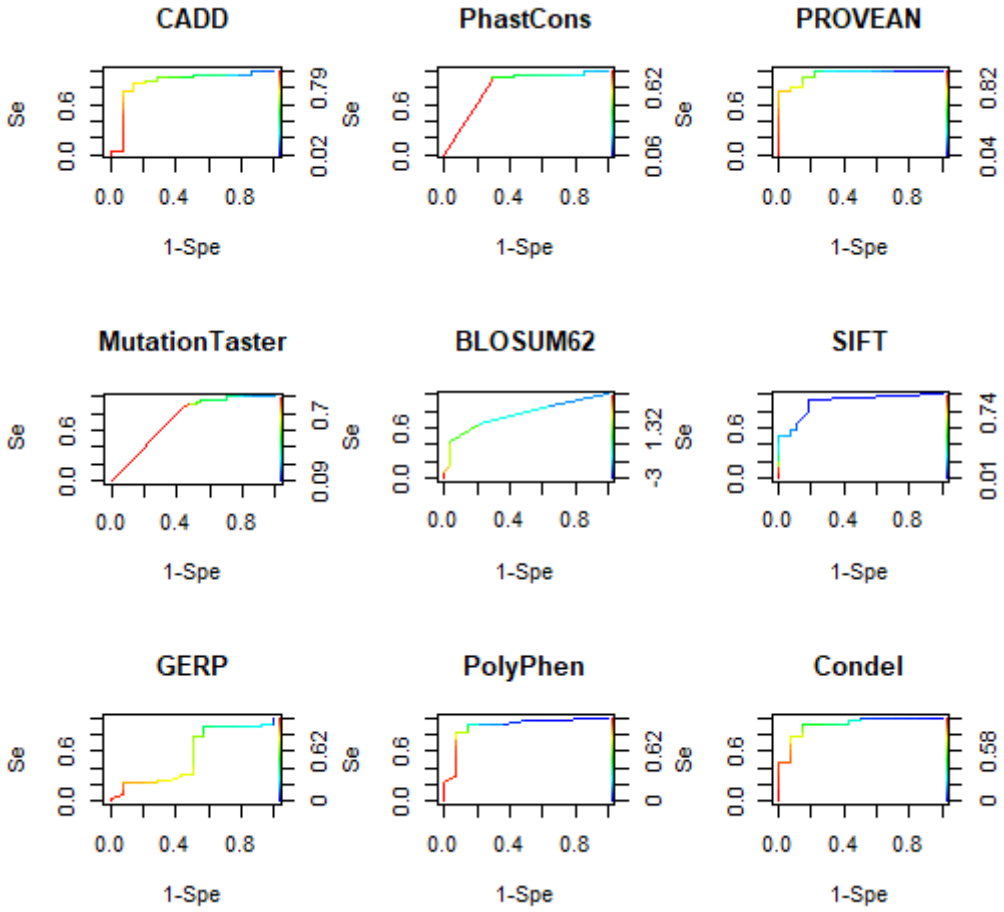

Figure 33. Courbes ROC obtenues pour les 9 outils de prédiction sur les variants étudiés issus du set RP.

Logiciel	AUC	Se	Spe	VPP	VPN
Condel	0,92	0,89 (0,78-1)	0,86 (0,67-1)	0,93 (0,83-1)	0,80 (0,59-1)
SIFT	0,89	0,82 (0,68-0,96)	0,79 (0,57-1)	0,88 (0,77-1)	0,69 (0,44-0,93)
PolyPhen	0,90	0,89 (0,78-1)	0,86 (0,67-1)	0,93 (0,83-1)	0,80 (0,59-1)
GERP	0,55	0,71 (0,55-0,88)	0,50 (0,24-0,76)	0,74 (0,58-0,9)	0,47 (0,20-0,72)
BLOSUM62	0,75	0,96 (0,90-1)	0,43 (0,17-0,69)	0,77 (0,62-0,93)	0,86 (0,67-1)
PROVEAN	0,96	0,89 (0,78-1)	0,86 (0,67-1)	0,93 (0,83-1)	0,80 (0,59-1)
MutationTaster	0,79	1 (1-1)	0,43 (0,17-0,69)	0,78 (0,62-0,93)	1 (1-1)
PhastCons	0,82	0,89 (0,78-1)	0,71 (0,48-0,95)	0,86 (0,73-0,99)	0,77 (0,55-0,99)
CADD	0,86	0,89 (0,68-0,96)	0,86 (0,67-1)	0,92 (0,82-1)	0,71 (0,47-0,94)

Tableau 10. AUC (aire sous la courbe) des courbes ROC, performances des logiciels de prédiction et intervalles de confiance à 95% déterminés à partir des variants issus du set Rétinites Pigmentaires. Se : Sensibilité, Spe : Spécificité, VPN : Valeur Prédictive Négative, VPP : Valeur Prédictive Positive.

2.2.2. Performances croisées

L'arbre de classement obtenu permet de reclasser l'ensemble des variants grâce à l'utilisation conjointe des logiciels PROVEAN et SIFT (Figure 34). PROVEAN semble donc être le logiciel qui présente le plus d'intérêt dans la thématique RP, que ce soit au niveau de ces performances individuelles ou également dans l'utilisation combiné aux autres algorithmes (SIFT ici).

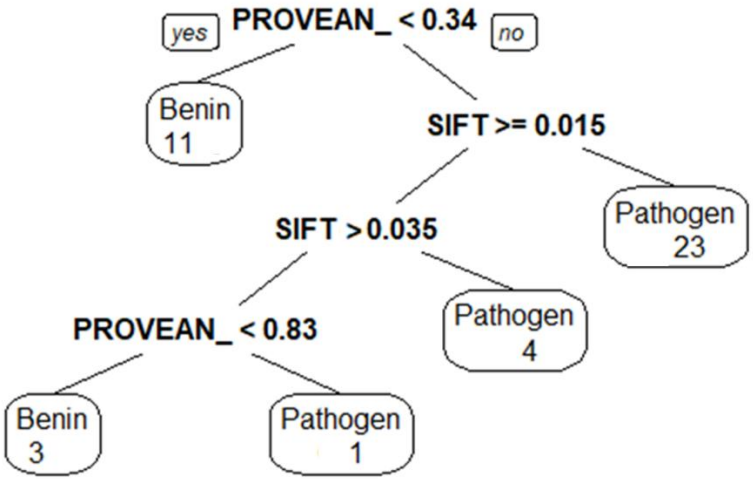

Figure 34. Arbre de classement obtenu par la méthode CART sur les variants étudiés issus du set Rétinites Pigmentaires.

2.3. Variants issus de la base déficience intellectuelle

2.3.1. Performances individuelles des logiciels

Les courbes ROC obtenues pour les 9 outils montrent des performances qui ne sont pas dues au hasard (AUC>0,5) (Figure 35). Les valeurs de performances ainsi que leur intervalle de confiance à 95% sont présentés dans le tableau 11. Les 2 logiciels présentant les meilleures performances individuelles globales sont Condel et PolyPhen.

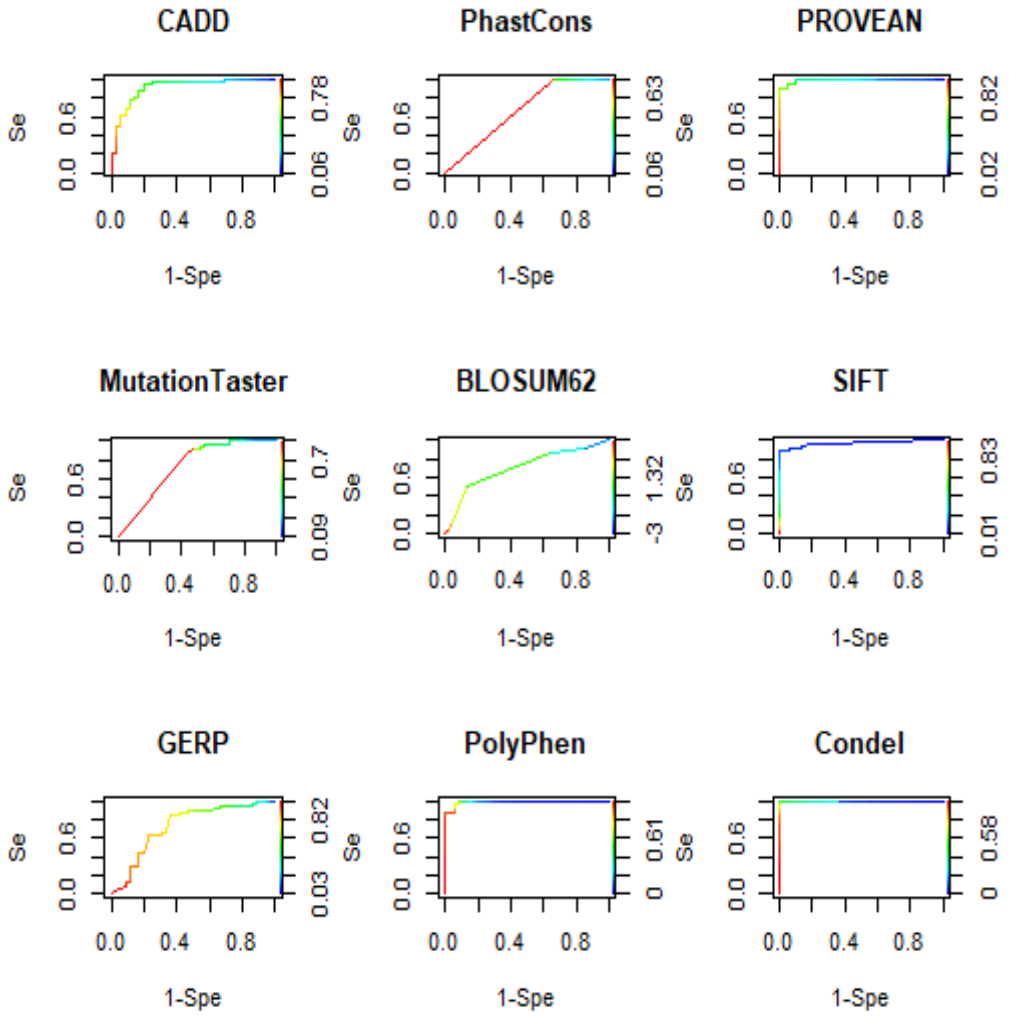

Figure 35. Courbes ROC obtenues pour les 9 outils de prédiction sur les variants étudiés issus du set Déficience Intellectuelle.

Logiciel	AUC	Se	Spe	VPP	VPN
Condel	0,99	0,99 (0,96-1)	1 (1-1)	1 (1-1)	0,97 (0,92-1)
SIFT	0,97	0,93 (0,87-0,99)	0,89 (0,79-0,98)	0,78 (0,68-0,88)	0,97 (0,92-1)
PolyPhen	0,99	0,96 (0,91-1)	0,94 (0,87-1)	0,97 (0,93-1)	0,92 (0,83-1)
GERP	0,73	0,66 (0,55-0,77)	0,67 (0,52-0,81)	0,79 (0,70-0,89)	0,50 (0,34-0,66)
BLOSUM62	0,70	0,96 (0,91-1)	0,11 (0,01-0,21)	0,68 (0,57-0,79)	0,57 (0,42-0,73)
PROVEAN	0,98	0,89 (0,82-0,97)	0,95 (0,88-1)	0,94 (0,89-1)	0,90 (0,81-0,99)
MutationTaster	0,78	1 (1-1)	0,25 (0,11-0,39)	0,72 (0,62-0,83)	1 (1-1)
PhastCons	0,67	0,99 (0,96-1)	0,36 (0,21-0,51)	0,75 (0,65-0,85)	0,93 (0,84-1)
CADD	0,92	0,93 (0,87-0,99)	0,81 (0,68-0,92)	0,90 (0,83-0,97)	0,85 (0,74-0,96)

Tableau 11. AUC (aire sous la courbe) des courbes ROC, performances des logiciels de prédiction et intervalles de confiance à 95% déterminés à partir des variants issus du set Déficience Intellectuelle. Se : Sensibilité, Spe : Spécificité, VPN : Valeur Prédictive Négative, VPP : Valeur Prédictive Positive.

2.3.2. Performances croisées

L'arbre de classement obtenu permet de reclasser la quasi-totalité des variants grâce à l'utilisation de Condel (Figure 36). Ce logiciel semble donc être celui qui présente le plus d'intérêt dans la thématique DI, que ce soit au niveau de ces performances individuelles mais également dans l'utilisation combinée.

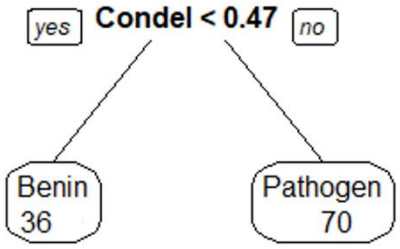

Figure 36. Arbre de classement obtenu par la méthode CART sur les variants étudiés issus du set Déficience Intellectuelle.

2.4. Comparaison selon les thématiques

La comparaison des performances des différents logiciels montre des performances variables en fonction des thématiques (Figure 37). Concernant Condel, les performances sont assez stables en fonction de la thématique. Nous observons tout de même de meilleures performances avec le set de variants de DI par rapport aux deux autres. Ces similitudes de variations sont retrouvées avec les logiciels Polyphen et PROVEAN. CADD montre une valeur moindre de VPN dans la thématique RP. Les prédictions de MutationTaster montrent dans tous les cas une sensibilité et une VPN de 100%, cependant les valeurs de spécificité et de VPP diffèrent entre les thématiques tout en restant médiocres (spécificité <50% quelle que soit la thématique), donc il est très probable qu'un variant soit bénin si cet outil le prédit non délétère. Les performances de SIFT sont assez variables selon la thématique, une bonne VPN ressort pour les thématiques DI et OD mais celle-ci est moyenne dans la thématique RP. A l'inverse la VPP est meilleure dans la thématique RP que pour les autres. Il y a une inversion du rapport VPP/VPN en fonction de la thématique pour le logiciel SIFT.

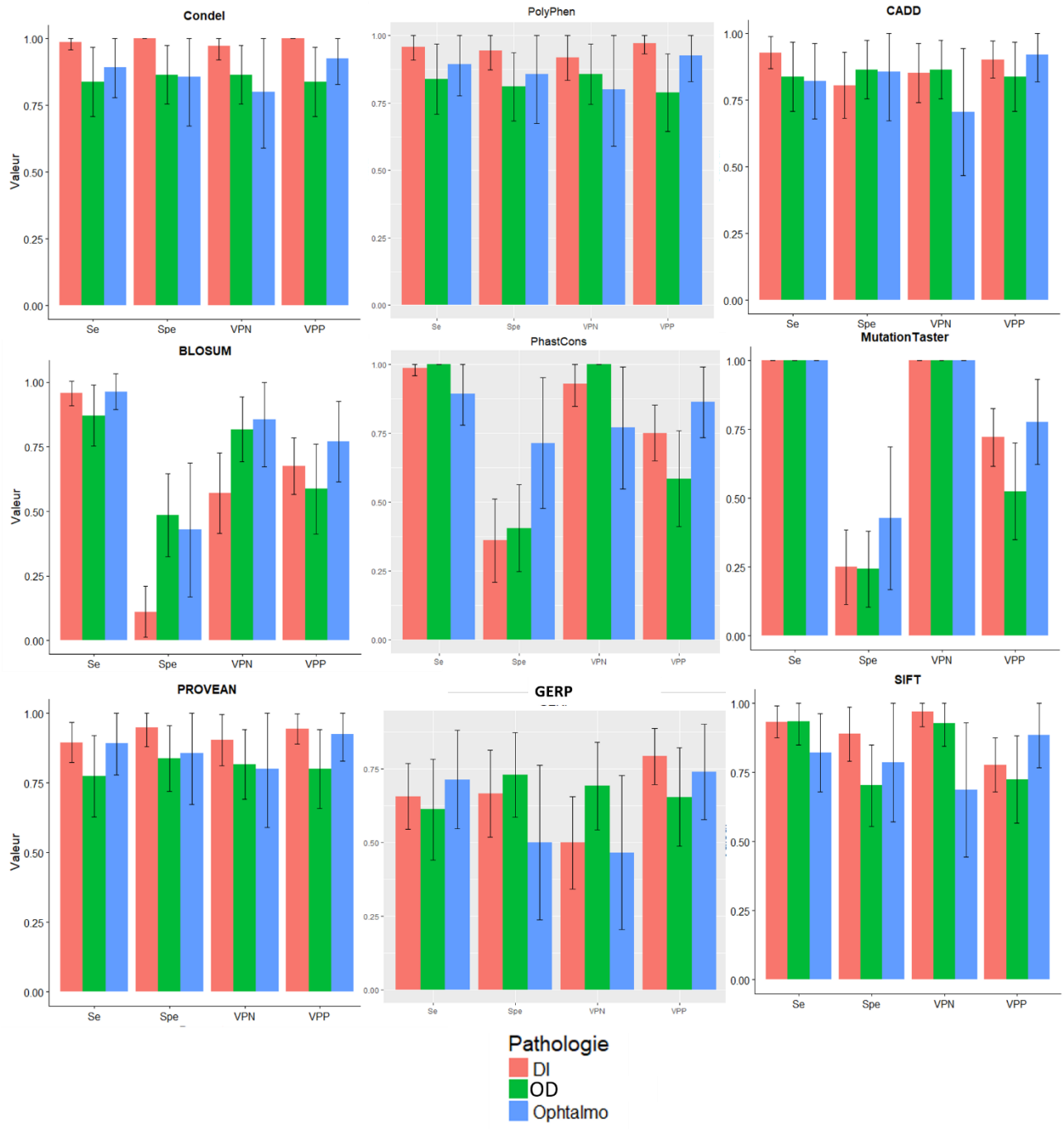

Figure 37. Comparaison des performances de chaque logiciel en fonction des thématiques, DI (déficience intellectuelle), Ophtalmo (rétinites pigmentaires) et OD (Oncogénétique Digestive). Les performances sont représentées avec leur intervalle de confiance à 95%. Se : Sensibilité, Spe : Spécificité, VPN : Valeur Prédictive Négative, VPP : Valeur Prédictive Positive.

3. Discussion

Nous avons pu tester les outils de prédictions *in silico* de l'effet des mutations faux sens sur 3 sets de données. Ces données diffèrent à la fois par la nature des gènes impliqués, les fonctions des protéines codées et par les critères de sélection et de classification des variants, avec une implication plus ou moins importante des résultats des logiciels de prédiction *in silico* dans le classement final du variant en délétère ou bénin. En effet, plus le diagnostic génotypique d'une maladie est ancien, les gènes impliqués connus et les relations entre phénotype et génotypes étudiées, plus les variants seront documentés.

Les performances de chaque logiciel sont comparées pour chaque thématique dans le but de s'assurer qu'il n'y ait pas de discordances ou biais selon la provenance des variants analysés. À la vue de nos résultats, il semble que la performance des logiciels soit dépendante de la maladie étudiée que ce soit au niveau des performances individuelles ou des résultats des arbres de classement. Il serait donc intéressant que chaque thématique, via son réseau, puisse réaliser l'évaluation des outils in silico sur son propre panel de gênes afin de déterminer la meilleure combinaison de logiciels possible, mais également de connaitre les performances propres de chaque logiciel. Ces différences peuvent être liées aux différences de fonction des protéines codées par les gènes étudiés. Néanmoins, une implication plus ou moins importante des résultats des logiciels de prédiction dans le classement final du variant en délétère ou bénin ne peut être éliminée.

Dans notre étude, dans un seul cas de figure, la quasi-totalité des variants a pu être correctement classée par le logiciel Condel, mais ceci n'est valable que sur les variants issus de la thématique déficience intellectuelle. Dans cette thématique, les variants sélectionnés ont été majoritairement classés dans la littérature et sont issus des bases de données. Il est donc possible que nous ayons introduit un biais à ce niveau concernant ce logiciel dont la mise au point de l'algorithme initial a été réalisé à partir de variants issus des bases de données et de la littérature (46). Cependant ce logiciel présente également de bonnes performances globales avec les autres sets de variants analysés.

Les résultats montrent que les performances annoncées par les développeurs ne sont pas toujours fidèles à ce que nous retrouvons avec nos sets de données. Par exemple la spécificité de MutationTaster est toujours très inférieure à celle annoncée puisqu'au maximum nous retrouvons une spécificité à 40 % contre plus de 80% annoncée. Il peut

donc être intéressant de réaliser une évaluation de ces performances avant utilisation d'un logiciel.

Nous mettons en avant les performances propres de chaque outil étudié. De cette façon, les logiciels pourraient être choisis en fonction du variant étudié. En effet, si d'autres arguments tels que la ségrégation ou la confrontation avec la clinique ne sont pas en faveur de la pathogénicité du variant, il semble plus intéressant d'utiliser un outil de prédiction possédant une bonne valeur prédictive négative pour compléter l'argumentation. De même si les autres arguments sont en faveur de la pathogénicité du variant, il est alors plus intéressant d'utiliser un outil de prédiction possédant une bonne valeur prédictive positive pour confirmer la pathogénicité. La bonne utilisation de ces outils en fonction de leur performance propre peut être une aide précieuse à la classification des variants.

Comme dit précédemment, le mauvais classement d'un variant peut entrainer des conséquences importantes sur la prise en charge du patient. Il est par conséquent risqué d'accorder trop d'importance à leurs résultats. Les logiciels de prédiction ne font qu'ajouter un argument supplémentaire aux données cliniques, aux analyses de ségrégation et données de fréquence du variant. Dans certains cas de figures, ces logiciels ne permettront pas de classer le variant et le recours aux tests fonctionnels sera nécessaire.

III. Les tests fonctionnels

Les outils de prédiction de l'effet délétère des variants restent des outils *in silico* basés sur des algorithmes et ne représentent qu'une aide à l'interprétation des variants. D'autres outils sont donc nécessaires afin d'apporter des arguments supplémentaires pour classer les variants d'interprétation difficile. Nous aborderons dans la suite de ce manuscrit, les tests fonctionnels utilisés en pratiques au CHRU de Lille, mais également ceux qui sont en développement ainsi que les perspectives pouvant être envisagées. Les informations sont issues des entretiens réalisés avec les biologistes du pôle de Biologie du CHRU de Lille concernant l'intérêt des tests fonctionnels en complément du séquençage massif dans leurs différentes thématiques.

Le test fonctionnel permet d'évaluer l'effet qu'un variant génétique aura sur la transcription (étude sur l'ARNm), sur la traduction ou les modifications post-traductionnelles (étude protéique) ou les conséquences sur les voies métaboliques impliquant la protéine

synthétisée (étude des métabolites) (Figure 38). Pour observer ces effets, il est possible dans certains cas de réaliser l'étude directement chez le patient (*in vivo*), mais la plupart du temps cela sera impossible. Il conviendra alors de « mimer » l'effet de la mutation dans un environnement contrôlé *in vitro*.

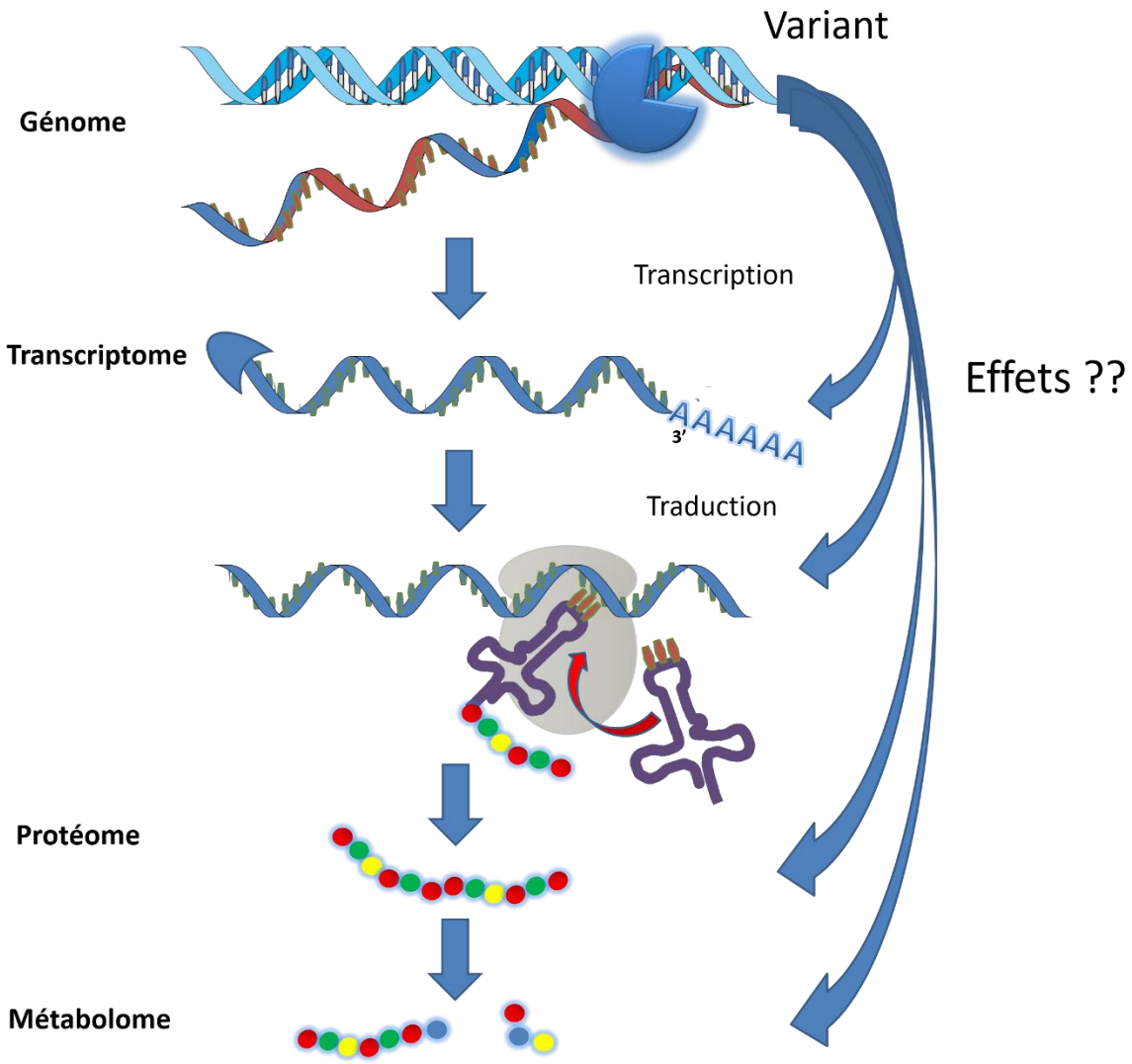

Figure 38. Les tests fonctionnels ont pour but d'évaluer les conséquences du variant depuis la transcription jusqu'aux modifications post-traductionnelles ainsi que sur la fonction biologique de la protéine traduite.

1. Evaluation de l'effet d'un variant sur les transcrits

L'épissage des pré-ARNm est une étape qui peut engendrer divers transcrits finaux et augmenter la variabilité protéique. Un épissage anormal peut entrainer des modifications importantes sur la protéine en aval. Les variants d'épissage peuvent toucher les sites

consensus donneurs ou accepteurs d'épissage, les sites de branchement ou encore les sites régulateurs (enhancer et silencer). Il est donc intéressant de caractériser l'impact du variant sur l'épissage (59) (Figure 39).

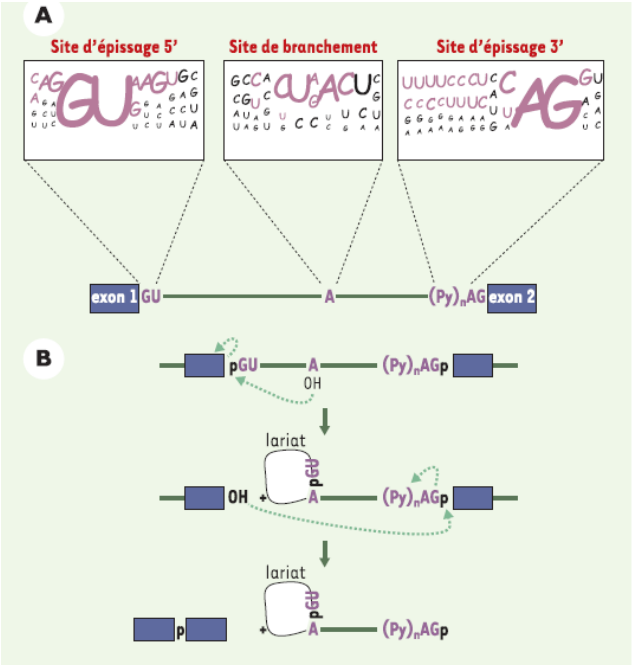

Figure 39 Représentation de l'épissage des pré-ARNm (59).

1.1. Tests réalisés en pratique au Centre de Biologie Pathologie

Un variant peut entrainer un effet sur l'épissage pouvant aboutir soit à la synthèse d'un transcrit anormal (exclusion d'exon, inclusion de tout ou partie d'intron) soit à une absence de transcrit par mécanisme NMD (*Nonsense Mediated mRNA Decay*). Ce mécanisme élimine les ARNm qui comportent un codon stop prématuré. Pour évaluer l'impact potentiel d'un variant sur l'épissage, il est possible de séquencer l'ARNm extrait à partir des globules blancs du patient, si le gène est exprimé dans ces cellules. La technique de RT-PCR, utilisant une rétrotranscriptase, permet la synthèse d'un ADNc (complémentaire) à partir de l'ARNm. L'ADNc est ensuite amplifié et séquencé. Le séquençage des transcrits est couramment réalisé en diagnostic pour les nouveaux variants, en particulier lorsque les logiciels de prédiction sont en faveur d'un effet du variant sur l'épissage.

Il est également possible de quantifier par une technique de PCR quantitative les transcrits mutés par rapport aux transcrits sauvages et de comparer l'expression de l'allèle porteur du variant avec celle de l'allèle sauvage.

1.2. Tests en cours de développement

Pour analyser les conséquences d'un variant sur l'épissage, la technique du minigène est une méthode *ex* vivo qui consiste à fabriquer un vecteur constitué d'un promoteur ubiquitaire suivie par l'exon que l'on souhaite étudier et ses séquences introniques adjacentes ainsi qu'un exon après le promoteur afin que l'épissage puisse se faire. Le tout est terminé par un exon portant un signal de polyadénylation pour stabiliser les transcrits (Figure 40). Cette construction est ensuite transfectée dans une lignée cellulaire, les ARN sont extraits plusieurs heures après la transfection et analysés par RT-PCR. Une version de la construction contient la version mutée de l'exon et une version contient la séquence sauvage qui sert de contrôle. Il est donc possible à ce niveau d'étudier l'impact de la mutation sur la l'épissage de l'exon par comparaison de la construction avec le variant avec la construction minigène contenant la séquence sauvage correspondante (60).

Par rapport au séquençage de l'ARN du patient, ce test d'épissage *ex vivo* présente l'avantage de permettre l'étude des allèles mutés et sauvages séparément. De plus, la construction est réalisée à partir de l'ADN du patient : elle permet donc l'étude des gènes qui ne sont pas exprimés dans les lymphocytes. Elle est cependant plus longue et plus complexe à mettre en œuvre que la RT-PCR sur lymphocytes.

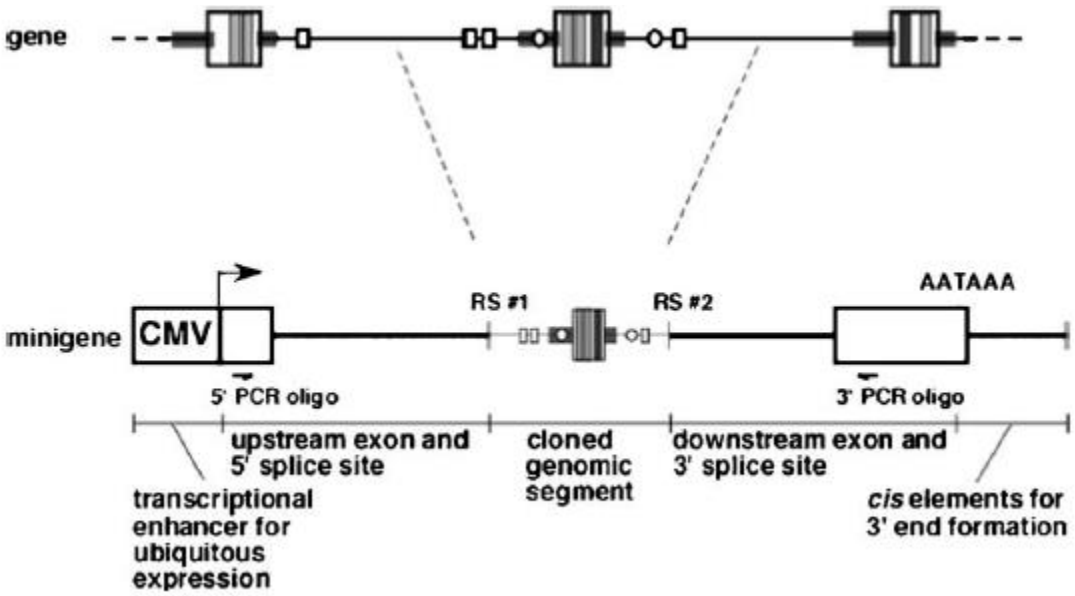

Figure 40 Principe de construction d'un minigène. RS : Site de restriction (61).

Cette technologie est aujourd'hui développée au CHRU de Lille dans le cadre de la thématique syndrome de Lynch, mais son utilisation devrait s'étendre à d'autres

thématiques telle que la maladie de Stargardt qui implique le gène *ABCA4* dont une part non négligeable de variants sont situés sur des sites d'épissage non canoniques. Il pourrait être également intéressant de la développer dans d'autres pathologies afin de s'assurer que les variants étudiés n'impactent pas la transcription avant de développer des outils fonctionnels protéiques nécessitant de longues mises au point.

2. Evaluation de l'effet d'un variant sur la protéine

La caractérisation fonctionnelle biochimique en parallèle du séquençage permet de conforter l'implication d'un variant identifié sur le gène codant la protéine impliquée. Nous illustrons ce type d'explorations par une liste d'exemples non exhaustives de tests fonctionnels évoqués après discussion avec les biologistes impliqués dans le séquençage au sein des Unités Fonctionnelles (UF) du Centre de Biologie et Pathologie de Lille. Nous abordons d'une part les tests fonctionnels intégrés dans les laboratoires de diagnostic et d'autre part les tests fonctionnels réalisés dans les laboratoires de recherche et qui ne sont pas intégrés en diagnostic.

2.1. Tests intégrés au laboratoire de diagnostic

2.1.1. Explorations *in vivo* chez le patient

2.1.1.1. Implication dans une voie métabolique

Si le gène incriminé code une protéine impliquée dans une voie biochimique bien connue et dont l'implication dans la maladie est bien établie, il est possible de réaliser directement les analyses fonctionnelles à partir d'un prélèvement réalisé chez le patient. Il peut s'agir d'une analyse quantitative (dosage protéique) ou qualitative (mesure de l'activité métabolique).

2.1.1.1.1. Métabolisme et déficience intellectuelle

La déficience intellectuelle (DI) est définie comme une efficience intellectuelle inférieure à un score de QI de 70. La DI est retrouvée dans de très nombreuses pathologies très hétérogènes cliniquement et concerne 2,5% de la population générale. On distingue les DI syndromiques, c'est-à-dire associées à d'autres symptômes, et les DI isolées. Les DI peuvent être la conséquence de maladies héréditaires du métabolisme. Une collaboration entre l'UF « Métabolisme Général et Hormonal - Maladies rares » et l'UF « Plateforme de puces à ADN » permet dans certains cas de confronter les données de métabolomique ciblée aux données de séquençage.

Un déficit enzymatique dans une voie métabolique entraine une accumulation des métabolites ou substrats en amont de l'enzyme déficitaire et une diminution des produits en aval. L'enzyme en cause pourra alors être identifiée et son activité pourra être mesurée afin de confirmer le déficit et d'établir le diagnostic. C'est le cas pas exemple du déficit en pyruvate déshydrogénase (PDHC) qui provoque des encéphalomyopathies infantiles. La PDHC est codé par le gène *PDH1A* (62) et est impliqués dans le contrôle du métabolisme oxydatif des substrats énergétiques . La découverte d'un variant dans ce gène corrélé à une accumulation des métabolites en amont de la PDHC, principalement le pyruvate, conforte alors l'implication du variant dans la pathologie du patient. Il en est de même pour les variants retrouvés sur le gène *PRODH* qui code pour la proline déshydrogénase impliquée dans le métabolisme des acides aminés. L'identification d'un variant de ce gène, une accumulation des métabolites en amont (proline) et un déficit des métabolites en aval (pyrroline 5 carboxylate) conforte l'implication pathologique du variant retrouvé (Figure 41) (63).

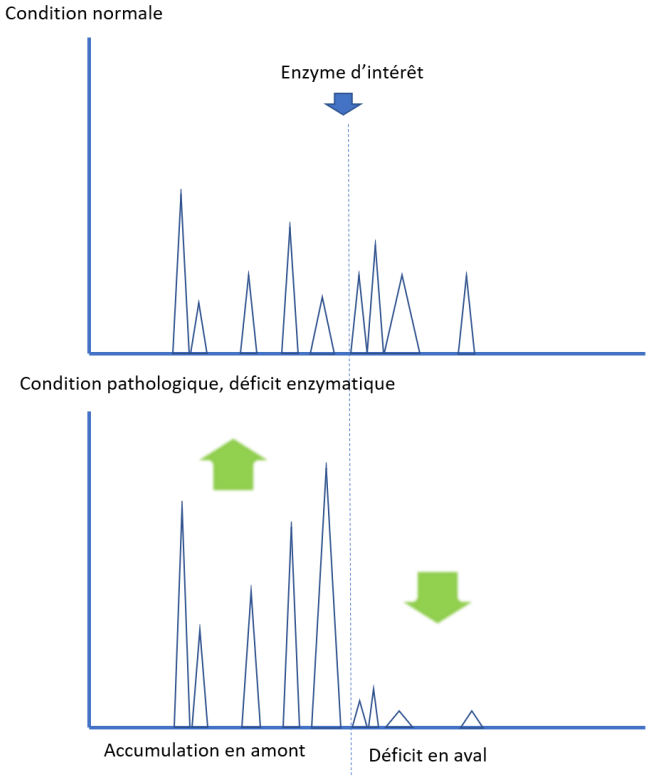

Figure 41. Représentation schématique de deux chromatogrammes d'étude d'une voie métabolique. Impact du déficit impliqué dans la voie métabolique (représenté par la flèche bleue) sur la quantification des métabolites impliqués (pics bleus). Condition normale en haut et condition pathologique (déficit de l'enzyme d'intérêt) en bas. La flèche bleue représente la limite entre produit et substrat.

2.1.1.1.2. Déficit en alpha1antitrypsine constitutionnel

Le déficit en alpha1antitrypsine (A1AT) constitutionnel est une pathologie considérée comme autosomique récessive. Cette maladie est étudiée dans l'UF « Métabolisme Général et Hormonal - Maladies rares ». Les principales manifestations cliniques de cette maladie sont pulmonaires (emphysème) et hépatiques (cirrhose). L'A1AT appartient à la superfamille des serpines, inhibiteurs de protéase à sérine. Son enzyme cible est l'élastase leucocytaire libérée par dégranulation des polynucléaires neutrophiles lors des réactions inflammatoires. Il s'agit du principal inhibiteur de protéase dans le sang. Elle contribue à la protection du parenchyme pulmonaire et participe à la régulation de la réponse inflammatoire. Le gène codant pour cette protéine est *SERPINA1* et présente un polymorphisme très important (64).

Dans le cadre du diagnostic de déficit en A1AT, les investigations biologiques débutent généralement par les explorations fonctionnelles : le dosage sérique de l'A1AT et la mesure de l'activité anti-élastasique. Le phénotypage de l'A1AT par isoélectrofocalisation en gel d'agarose permet l'identification des variants les plus courants comme le variant PI M, des variants déficitaires les plus fréquents PI S (p.Glu264Val) et PI Z(p.Glu342Lys), mais aussi des variants plus rares comme PII, PI F, PI E et PI P. Les explorations génétiques sont réalisées dans 10% des cas uniquement à l'issue de la détermination phénotypique afin de confirmer ou infirmer l'hypothèse diagnostique initiale. La recherche de mutation s'effectue par séquençage complet du gène *SERPINA1* par technique Sanger ou par NGS (65). Le séquençage permet non seulement de caractériser les variants fréquents déficitaires ou non, mais également les variants plus rares, les variants null, voire des variants non encore décrits. Le diagnostic génotypique permet d'affirmer le caractère constitutionnel de l'anomalie, ce qui a une très grande importance dans la prise en charge du patient. En effet le traitement substitutif (Alfalastin®) obtenu par fractionnement plasmatique a une indication stricte limitée au déficit constitutionnel avéré, il est donc essentiel de caractériser les variants identifiés en NGS sur les gènes causaux. L'analyse fonctionnelle réalisée en parallèle directement chez le patient est donc complémentaire et va aider à l'interprétation des données de séquençage obtenues par NGS et permet une caractérisation robuste des variants identifiés.

2.1.2. Modèles cellulaires

2.1.2.1. Troubles constitutionnelles de l'hémostase

L'hémophilie est une maladie hémorragique constitutionnelle liée à un déficit en facteur de la coagulation : facteur VIII pour l'hémophilie A, facteur IX pour l'hémophilie B. C'est une maladie héréditaire dont le mode de transmission est récessif lié au chromosome X : seuls les garçons sont atteints (66,67). Le diagnostic moléculaire par séquençage à haut débit a été mis en place en 2015 au CHRU de Lille dans l'UF « Hémostase Spécialisée et Moléculaire » (68). La caractérisation génétique de la pathologie permet d'adapter à la fois la prise en charge thérapeutique et de délivrer un conseil génétique aux familles. Lors de la découverte d'un variant d'un patient hémophile, la première étape est de vérifier si celui-ci a déjà été rapporté dans les bases de données spécifiques de variants pour l'hémophilie (http://factorviii-db.org/). Si un variant n'est pas répertorié ou difficilement classable, une collaboration avec une équipe lyonnaise propose un test fonctionnel pour le caractériser. Il s'agit d'un protocole utilisant de la mutagénèse dirigée des gènes *F8* et *F9* sur une lignée cellulaire COS-1 exprimant les facteurs VIII et IX. Le niveau d'expression des facteurs dans les cellules exprimant la version mutée du gène (mt) est comparée à celui des cellules non mutées (wt) (Figure 42) (67). Cela permet d'observer l'effet du variant moléculaire d'une part au niveau qualitatif (activité du facteur VIII) et d'autre part au niveau quantitatif (dosage du facteur VIII).

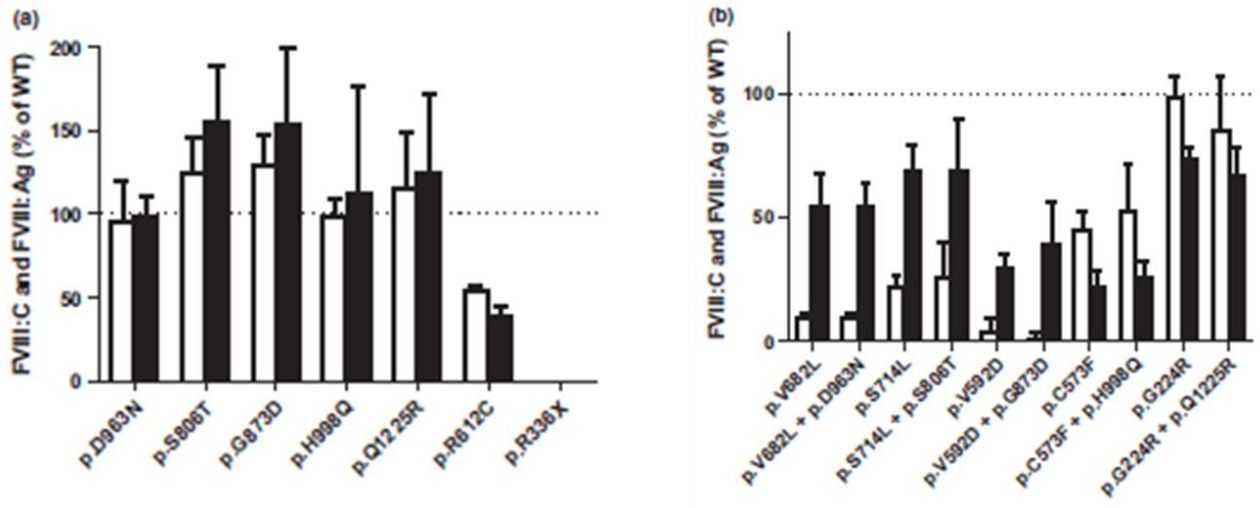

Figure 42. Niveau d'expression du facteur VIII sur des cellules COS-1 muté par rapport aux cellules non mutées (WT). En blanc, les données qualitatives (activité du facteur VIII) et en noir le dosage quantitatif du facteur VIII (66).

2.1.2.2. Syndrome de Lynch

Le syndrome de Lynch est la forme la plus fréquente de prédisposition héréditaire aux cancers colorectaux. C'est une maladie à transmission autosomique dominante. L'étude génétique de cette pathologie est réalisée au sein de l'UF « Oncogénétique moléculaire ». Elle est liée à une mutation d'un des gènes codant des protéines participant au système de réparation des mésappariements de l'ADN *MisMatch Repair* (MMR). Les gènes impliqués sont *MLH1, MLH2, MSH6* et *PMS2 ;* la maladie est caractérisé par un défaut de réparation conduisant à un risque accru de développer un cancer notamment colo-rectal (69). Deux mutations délétères dans les cellules coliques (le 2^{e} événement somatique) entrainent un défaut de réparation et l'accumulation de variations nucléotidiques dans la cellule, dont des mutations dans des oncogènes ou des gènes suppresseurs de tumeur, ce qui va entrainer le développement de la tumeur. L'identification d'une mutation causale permet d'affirmer le diagnostic de syndrome de Lynch et d'adapter la surveillance du patient. En effet dans le syndrome de Lynch, la transformation d'un adénome en adénocarcinome est accélérée (de l'ordre de 2 à 3 ans contre 8 à 10 ans dans la population générale), ce qui justifie de réaliser une coloscopie tous les 2 ans (70). De nombreuses mutations liées au syndrome de Lynch sont tronquantes : l'implication dans la pathologie est alors considérée comme évidente. Cependant une part non négligeable des variations identifiées sont des variations faux-sens d'interprétation plus délicate. Certaines de ces variations ont déjà été décrites précédemment et ont été classées au niveau international ou national. Au laboratoire, dans le cas d'un travail de thèse, un test fonctionnel pour aider à la caractérisation des variants faux sens de *MLH1* et *MSH2* a été mis au point (71).

La séquence codante du gène sauvage est clonée dans un vecteur d'expression. Le variant d'intérêt est ensuite créé par mutagenèse dirigée. Après une transformation des bactéries compétentes, l'ADN plasmidique est extrait, puis la séquence de l'insert est vérifiée par séquençage. L'ADN plasmidique est introduit dans des cellules déficientes pour le gène étudié par une étape de transfection cellulaire. Parallèlement, des cellules sont transfectées dans les mêmes conditions avec la séquence sauvage du gène étudié. Les protéines sont extraites et analysées par Western Blot à différents temps après la transfection. La quantité de protéine produite est comparée entre la séquence avec le variant d'intérêt et la séquence sauvage, au même temps de transfection (Figure 43).

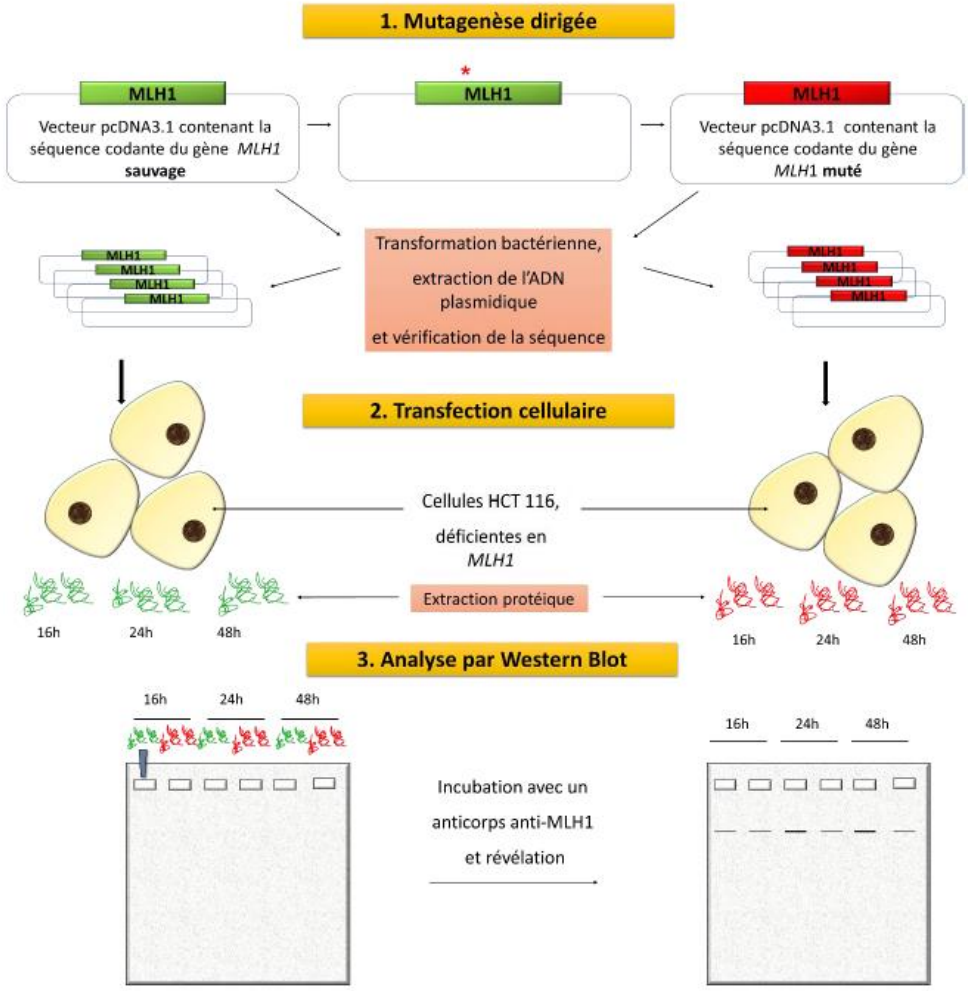

Figure 43. Principe du test fonctionnel pour les gènes *MLH1* et *MSH2* impliqués dans le syndrome de Lynch (71).

Ce test apporte un argument supplémentaire sur l'implication potentielle ou non de variant dans la pathologie. Dans le cadre de ce travail, 6 variants de *MLH1* et 12 variants de *MSH2* classés initialement VSI ou d'implication probable ont été étudiés et cette étude a permis d'apporter des arguments supplémentaires permettant d'aider à leur reclassement. Cependant, l'interprétation de certains résultats obtenus reste difficile et certains résultats sont parfois peu concordants avec les données disponibles (littérature et données clinico-biologiques). Ce test est une aide à l'interprétation des variants identifiés en NGS. La mise au point d'une telle technique impose cependant de disposer d'un modèle cellulaire dépourvu du gène fonctionnel, ce qui peut rendre le développement d'un tel test laborieux si le modèle n'existe pas. De plus un tel test n'évalue que l'expression protéique quantitative et ne prends pas en compte le côté fonctionnel des protéines. Un test complémentaire permettant d'évaluer la fonction de réparation MMR existe également.

2.2. Tests utilisés en laboratoire de recherche

2.2.1. Microscopie confocale et localisation subcellulaire

Lorsque la fonction de la protéine d'intérêt est peu ou mal connue, il peut être intéressant de réaliser une étude de sa localisation subcellulaire grâce à la microscopie confocale en utilisant un marqueur fluorescent tel que la GFP (*Green Fluorescent Protein*). Une mutagénèse dirigée permet d'exprimer le variant génétique dans un environnement *in vitro* (lignée cellulaire d'intérêt), auquel un marqueur fluorescent GFP est ajouté afin de localiser l'expression protéique cellulaire. Si on prend l'exemple d'une protéine codée qui a une localisation nucléaire, le marquage GFP est alors nucléaire. Si la protéine « mutée » a une localisation diffuse cytoplasmique (Figure 44), l'expérience montre que le variant génétique étudié entraine un défaut d'adressage cellulaire qui peut induire une perte de fonction. Ceci est un argument en faveur de l'implication du variant dans la pathologie du patient.

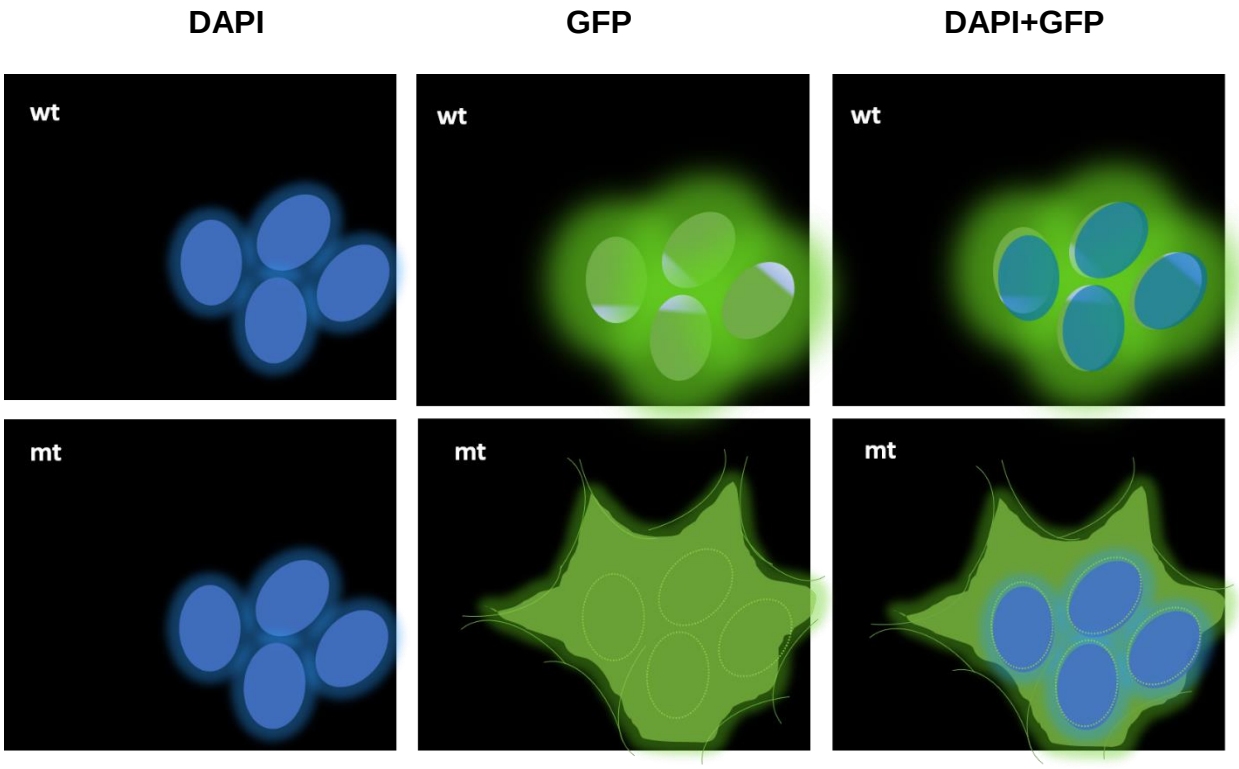

Figure 44. Représentation schématique du test fonctionnel utilisant la microscopie confocale sur une lignée de cellules Perte de la localisation cellulaire de la protéine en condition mt par rapport au wt. Noyaux marqués au DAPI (coloration bleue) et protéines d'intérêt marquées par la GFP (coloration verte). Condition non mutée : wt, condition mutée avec le variant étudiée : mt.

Des projets de recherche sont en cours afin de caractériser la pathogénicité des variants retrouvés dans les gènes de DI encore mal caractérisés. Dans ce contexte, chez un patient présentant une déficience intellectuelle une équipe du CHU de Tours a réalisé un test fonctionnel basé sur la localisation subcellulaire sur un variant faux sens d'un gène impliqué dans la DI qui était difficilement classable. Ce test a montré la pathogénicité du variant en lien avec la perte de la localisation nucléaire de la protéine mutée par rapport à la protéine sauvage.

2.2.2. Utilisation de la technologie CRISPR/Cas9.

La technologie innovante CRISPR/Cas9 présente un intérêt dans la mise en place de tests fonctionnels et notamment en vue de créer des lignées déficitaires si celles-ci n'existent pas. Ces cellules déficitaires pour le gène d'étude (cellules KO) peuvent être utilisées afin de mettre en évidence des cibles fonctionnelles par comparaison aux cellules normales (cellules wt) (Figure 45). Ces cibles peuvent être de différentes natures : il peut s'agir d'un

partenaire protéine, d'une cible transcriptionnelle, de conséquences protéiques telles que la méthylation d'autres cibles etc …

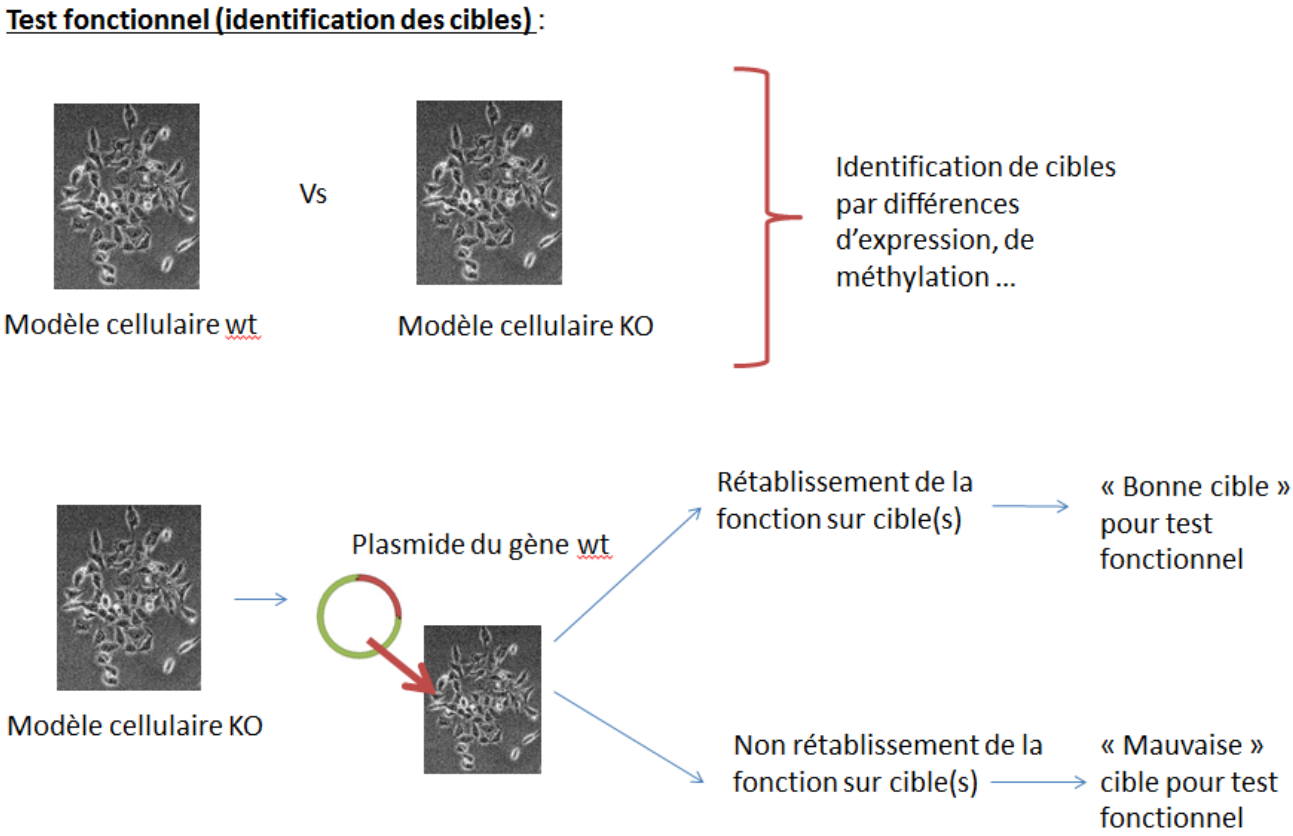

Figure 45. Identification de cibles fonctionnelles grâce à un modèle cellulaire dépourvu du gène à étudier (modèle KO) vs un modèle cellulaire sauvage (modèle cellulaire wt).

D'autre part, ces cellules KO peuvent être utilisées en vue de caractériser les variants difficile à interpréter. Une version mutée est alors exprimée dans ces cellules afin de statuer du caractère délétère ou non de la mutation. Les cibles fonctionnelles seraient alors perturbées seulement en cas de mutation délétère. (Figure 46).

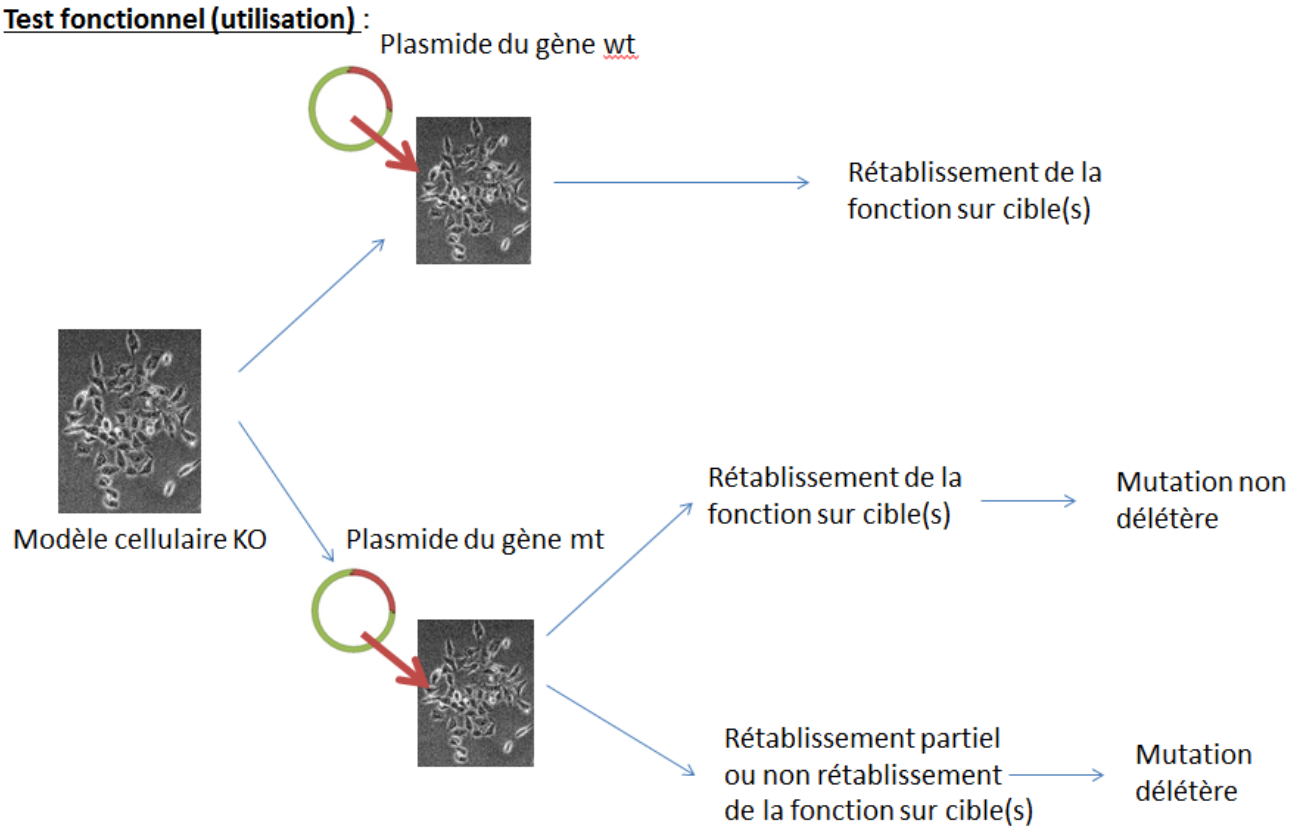

Figure 46. Utilisation du modèle cellulaire dépourvu du gène à étudier pour étudier l'impact de mutation par surexpression d'un plasmide porteur de la mutation (plasmide mt) ou d'un plasmide porteur du gène sauvage (plasmide wt).

Dans le cadre du projet de recherche réalisé dans le cadre de mon Master2, nous avons souhaité mettre au point de tels outils cellulaires en vue de développer un test fonctionnel pour caractériser les variants de 2 syndromes de déficience intellectuelle. Le syndrome de Wiedeman Steiner qui implique le gène *KMT2A* et le syndrome Kabuki qui implique les gènes *KMT2D et KDM6A* (72–75). Ces pathologies sont étudiées au CBP dans l'UF « Plateforme de puces à ADN ». Ces gènes codent des protéines impliquées dans la méthylation de l'histone H3 et interviennent dans le développement. Le but était de mettre au point des lignées cellulaires dépourvues des gènes incriminés telles que celles utilisées dans le test fonctionnel du syndrome de Lynch. Une des pistes envisagées était la mesure de la méthylation des lysines de l'histone H3 (73,76,77).

Pour cela nous avons voulu utiliser la technologie innovante CRISPR/Cas9 qui semblait rapide et facile d'utilisation (78) (Figure 47). Cette technologie est basée sur un système

de défense bactérien vis à vis des bactériophages (79). Le système CRISPR utilise la nucléase Cas9 qui, complexée à un ARN guide (ARNg) complémentaire de la séquence cible, clive l'ADN de façon spécifique en amont d'une séquence PAM (Protospacer Adjacent Motif) localisée sur l'ADN génomique. L'action de la nucléase Cas9 guidée par l'ARNg aboutit à une coupure double brin de l'ADN génomique au niveau de la séquence cible.

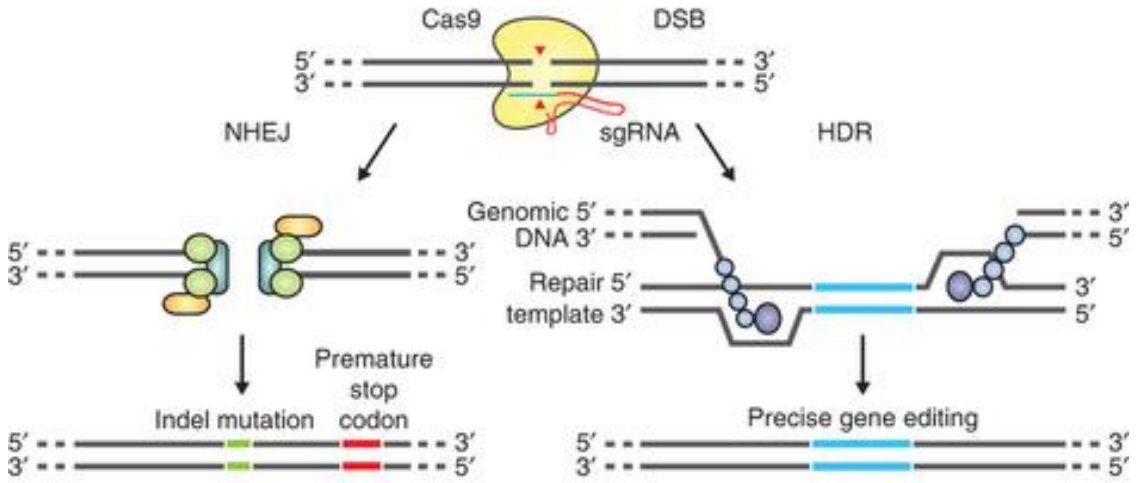

Figure 47 : Principe du CRISPR/Cas9. L'action de la nucléase Cas9 guidée par l'ARNg aboutit à une coupure double brin de l'ADN génomique au niveau de la séquence cible. Deux mécanismes de réparation peuvent se produire : NHEJ : jonction des extrémités non homologues, aboutissant au Knock-Out (KO) du gène cible ; HDR Recombinaison homologue : en présence de séquences homologues, insérées via un vecteur donneur, permettant de réaliser des Knock-Out/In, réparation ou délétion de gêne (78).

La littérature a montré que la technologie CRISPR/Cas9 pouvait générer beaucoup plus d'effets hors cibles (*off target*) que d'autres technologies (80). Il a été retrouvé jusqu'à 5 mésappariements pour un ARNg de 20 pb, d'où l'importance de choisir les ARNg les plus spécifiques possibles de la cible. Des solutions commerciales existent, proposant des plasmides qui combinent le ARNg, la Cas9 ainsi que des gènes rapporteurs, cependant elles ne proposent qu'au maximum 3 ARNg différents. Nous avons fait le choix de designer nous même les ARNg, d'une part afin de réduire cet effet *off-target* et d'autre part afin de choisir précisément notre cible. Ceci nous permet de réaliser une délétion d'au moins 200 pb dans chacun de nos gènes en choisissant des ARNg ciblant des exons proches pour créer une telle délétion (Figure 48).

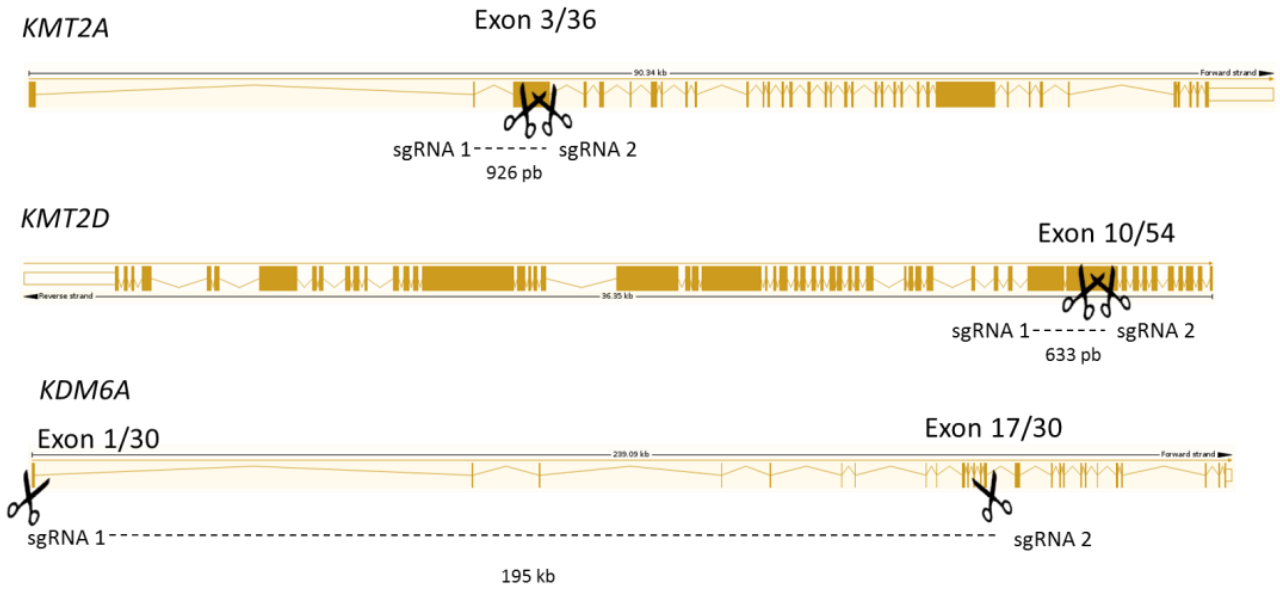

Figure 48 Localisation des cibles des ARNg sur la séquence génomique des différents gènes.

Ce travail nous a permis d'obtenir une lignée cellulaire HEK293T porteuse d'une délétion hétérozygote du gène *KDM6A*, mais nous ne sommes pas parvenus à obtenir de délétion pour les gènes *KMT2A* et *KMT2D*. Cette étude a requis plusieurs mois de mise au point, nécessitant à la fois des étapes bioinformatiques pour la sélection des cibles (ARNg), la construction de plasmides complexes, le clonage et la transfection dans des cellules à maintenir en culture. Le rendement de « mutation » étant encore très faible, il faut maintenir de nombreuses colonies de cellules en culture jusqu'à obtention du clone mutant souhaité. Ceci représente une charge importante de travail qui ne semble adapté, en l'état actuel avec une pratique en routine. En effet, l'étude de chaque gène nécessiterait une mise au point spécifique dû aux nombreux problèmes de spécificité et faible rendement du mécanisme de CRISPR/Cas9. Cependant, cette technologie n'est pas à proscrire totalement et pourrait avoir un intérêt futur lorsque sa robustesse permettra une utilisation nécessitant un temps inférieur de mise au point. Celle-ci a d'ailleurs déjà montré son intérêt dans l'utilisation de nombreux tests fonctionnels mais qui sont pour le moment réservés au domaine de la recherche (81–83).

2.2.3. Les IPSC

Le modèle IPSC (*Induced Pluripotent Stem* Cell) correspond à l'utilisation de cellules souches pluripotentes induites (iPS) qui sont créées artificiellement en laboratoire en « reprogrammant » les fibroblastes du patient (Figure 49). Ces cellules sont ensuite différenciées en cellules d'intérêt pour permettre l'étude *ex-vivo* des effets du variant du patient sur la transcription et sur la protéine codée (84).

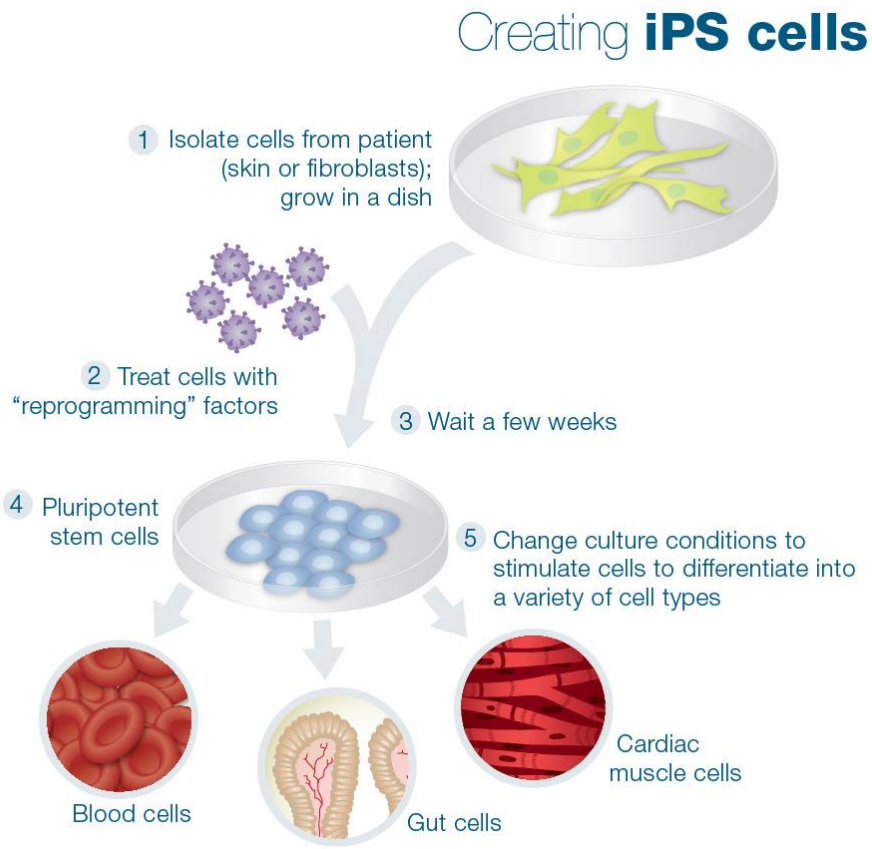

Figure 49. Principe général des IPSC (84)

L’approche par IPSC issus des fibroblastes de patient a permis de caractériser les variants des patients atteints de maladie de Stargardt une dystrophie maculaire héréditaire rare d'apparition précoce caractérisée par une perte progressive de la vision centrale. Cette maladie est étudiée au CBP dans l’UF «Génopathies et Pharmaco/Toxicogénétique » (85,86). Cette approche a été utilisée dans l’étude d’une autre dystrophie rétinienne, la choroïdérémie (CHM) qui implique le gène *REP1*, et a permis d’observer les différences entre cellules issus de patients malades par rapport à des cellules de patients sains (87). L'analyse qPCR montre une augmentation de l'expression (en unités relatives) des marqueurs génétiques de cellules rétiniennes typiques *RDH5, MERTK, ZO-1, TYR, BEST1, RLBP1* et *PAX6* dans des cellules IPSC sauvages (RPE WT) et cellules IPSC provenant d’un patient (RPE CHM6) par rapport aux fibroblastes (Fibro). Ceci montre la bonne différenciation génétique des cellules IPSC qui pourront être utilisées pour caractériser le variant d’intérêt (Figure 50). Différentes études ont été réalisées par la suite sur ces cellules IPSC sur les cibles connus de REP1 et ont montré l’implication du variant retrouvé en séquençage dans la pathologie par altération de la fonction de la protéine.

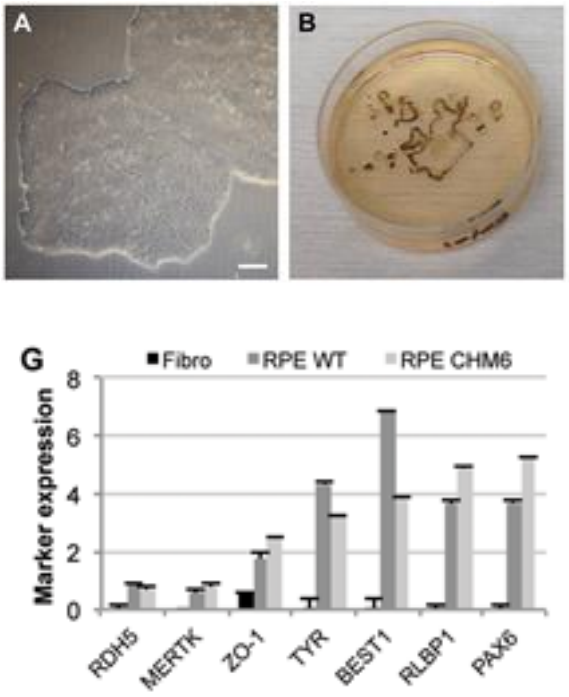

Figure 50. Génération, caractérisation et utilisation de cellules IPSC dans le cadre de la choroïdérémie. WT = condition sauvage, CHM6 = condition pathologique. (A) Colonies d'IPSC cultivées. (B) Cellules pigmentées apparaissant dans des plaques d'IPSC au cours du processus de différenciation spontanée. (G) l'analyse qPCR montre une augmentation de l'expression (en unités relatives) des marqueurs génétiques de cellules rétiniennes typiques *RDH5, MERTK, ZO-1, TYR, BEST1, RLBP1* et *PAX6* dans des cellules IPSC sauvages (barres gris foncé) et cellules IPSC provenant d'un patient (CHM6) (barres gris clair) par rapport aux fibroblastes (barres noires).

Cette stratégie a été utilisée dans d'autres thématiques, tel qu'en neurologie où il n'est pas possible de réaliser des études génotypiques directement sur neurones (88). Elle présente l'avantage de pouvoir travailler directement sur un prélèvement du patient afin de s'affranchir de tout biais lié à des constructions plasmidiques.

2.2.4. Etude des canaux

Lorsque le gène étudié code pour un canal ionique, il est possible d'étudier l'impact des variants grâce à des tests fonctionnels *in vitro* permettant l'étude des canaux. Le patch clamp est une technique électrophysiologique d'enregistrement des courants ioniques transitant à travers les membranes cellulaires. Les analyses de courant sont réalisées grâce à la technique du patch clamp sur cellule entière dans des modèles cellulaires par expression hétérologue des canaux sodiques en condition sauvage (WT) et mutée. Les

résultats sont ensuite comparés et permettent de déterminer si le variant génétique a ou non un impact sur la fonctionnalité du canal impliqué.

L'utilisation de cette technique a un intérêt dans l'exploration des variants retrouvés dans l'exploration génétique des surdités héréditaires, thématique étudié au CBP dans l'UF « Génopathies et Pharmaco/Toxicogénétique». La surdité est le déficit sensoriel le plus fréquent. 1/1000 à 1/700 enfants par an en France naissent avec une surdité profonde ou sévère. La grande majorité des surdités sont dites non-syndromiques ou isolées, c'est-à-dire que le déficit auditif est le seul signe clinique observé. Dans les pays développés, environ 60% des surdités précoces sont d'origine génétique. La plupart des surdités à la naissance sont des surdités de perception, par opposition aux surdités de transmission. 85 % des surdités isolées se transmettent selon un mode autosomique récessif (type DFNB), 10-15% selon un mode autosomique dominant (type DFNA), 1% selon un mode lié au chromosome X (type DFNX). La connexine GJB2 est impliquée dans 30 % des cas de surdité de perception non syndromique. Il est donc important de rechercher les mutations de ce gène en 1er intention (89–91). Les connexines sont des protéines transmembranaires de jonction, ou protéines de jonctions communicantes (Gap Junction proteins). Elles sont essentielles pour un grand nombre de processus physiologiques, tel que la dépolarisation coordonnée du muscle cardiaque. Les mutations affectant des gènes codant des connexines peuvent alors conduire à des anomalies fonctionnelles ou développementales. Lorsqu'un variant est retrouvé, il existe plusieurs bases de données propres à la pathologies permettant de le rechercher (92,93). De très nombreux autres gènes peuvent être impliqués dans les surdités d'origine génétique, certains de ces gènes codant des canaux. L'étude moléculaire est réalisée par séquençage haut débit. Le diagnostic précoce avant l'acquisition du langage est recommandé car il permet la prise en charge précoce de l'enfant et la proposition d'un appareillage ou d'un implant. Il est donc important de pouvoir caractériser les variants identifiés. Une des techniques d'étude fonctionnelle des mutations sur les connexines de surdité pourrait s'inspirer des études menées sur le gène *SCN5A* codant pour la sous-unité alpha du canal sodique cardiaque et impliqué dans le syndrome de Brugada. Les analyses du courant sodique ont été réalisées dans les conditions sauvages (WT) et mutée (R689H) grâce à la technique du patch clamp sur cellule entière, par expression hétérologue des canaux sodiques dans des cellules HEK293 (Figure 51). Cette étude a montré un effet délétère de la mutation d'intérêt qui empêche la dépolarisation du canal sodique (94).

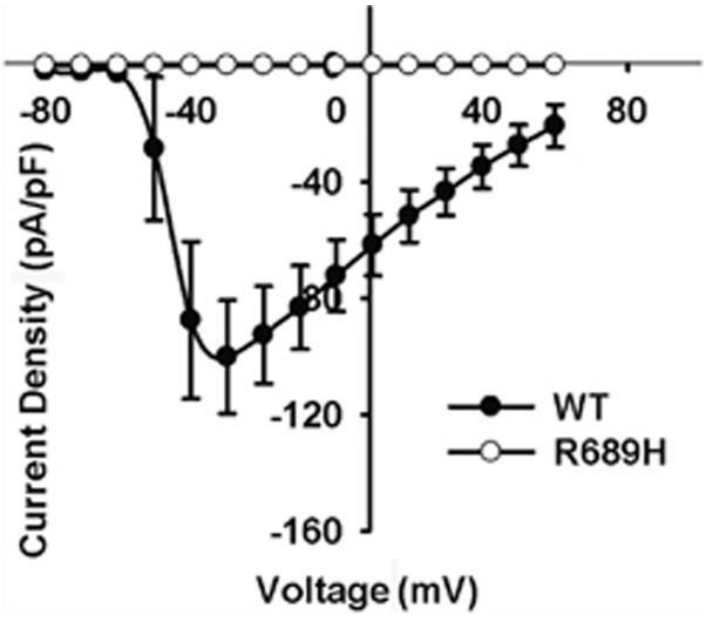

Figure 51. Dépolarisation du canal sodique sauvage WT et muté R689H exprimé dans des cellules HEK293 (94).

3. Discussion

Dans l'utilisation des tests fonctionnels protéiques, il peut être intéressant de s'assurer tout d'abord que la protéine puisse être traduite grâce à une étude préalable de la transcription. L'utilisation de la RT-PCR est possible sur les lymphocytes du patient mais nécessite que le gène d'intérêt soit exprimé dans ces cellules. Dans le cas contraire, le développement d'un test minigène permet d'étudier *ex vixo* l'impact d'un variant sur l'épissage mais nécessite des mises au point plus lourdes.

La confrontation des données biochimiques avec les résultats génétiques permet d'aider à caractériser l'implication d'un variant dans la pathologie. Des approches métaboliques ou par dosages qualitatifs et quantitatifs de la protéine d'intérêt peuvent dans certains cas être réalisées en parallèle du séquençage. Quand l'étude directe chez le patient n'est pas possible, le recours au test fonctionnel *in vitro* peut se révéler informatif. De tels tests vont toutefois requérir des mises au point lourdes et ne peuvent généralement pas être réalisés en routine. En effet, il n'existe pas de test fonctionnel universel et lorsque le gène est mal connu, de lourds travaux de caractérisation sont nécessaires. La collaboration avec des laboratoires de recherche est alors indispensable pour réaliser la mise au point de ses outils. Cependant lorsque le test fonctionnel *in vitro* est mis au point, un basculement de la technique au laboratoire de diagnostic est possible, à l'image de la technique du minigène.

Certaines approches fonctionnelles semblent avoir un grand intérêt tel que le modèle IPSC qui permet de travailler directement sur les cellules du patient et peut générer de nombreux types cellulaires, ou l'utilisation de CRISPR/Cas9 qui permet de cibler le génome à l'endroit souhaité. Mais ces techniques sont encore récentes et nécessitent beaucoup de mises au point et ne seront probablement pas utilisables en routine au laboratoire dans un avenir proche.

Pour le moment, il y a grande nécessité de collaboration avec des laboratoires de recherche qui pourront venir en aide au diagnostic en mettant au point des outils qui pourront être un jour utilisés en routine voir éventuellement le développement de méthodes universelles. La caractérisation des variants retrouvés en séquençage haut débit nécessiterait d'avoir des tests fonctionnels à haut débit.

IV. Conclusion

Les pratiques de séquençage actuelles tendent de plus en plus vers le séquençage à haut et très haut débit, l'objectif national à moyen terme étant un séquençage du génome ou de l'exome entier en diagnostic (Plan France Médecine Génomique 2025). Cependant, l'étude d'un génome génère plus de 3 millions de variants, il est donc essentiel de pouvoir identifier les variants pathogènes

Lorsqu'un variant n'est pas répertorié dans les bases de données, il convient d'utiliser diverses stratégies pour le classer. Pour les variants faux sens, une partie de cette approche consiste à utiliser des algorithmes bioinformatiques de prédiction de l'effet du variant sur la protéine codée. Les outils qui existent sont nombreux, il convient donc de les choisir et de les utiliser judicieusement. Ce travail a permis de comparer la performance de 9 outils de prédiction utilisés au laboratoire sur les variants identifiés dans 3 thématiques différentes. Cette étude des performances met en avant que les performances annoncées par les développeurs ne sont pas forcément retrouvées et que les performances peuvent varier en fonction de la thématique étudiée. Ce travail montre l'intérêt de réaliser une étude de performance avant utilisation de ces outils afin de choisir les outils les plus informatifs. Pour cela, il faut disposer de variants bien caractérisés issus des données de la thématique, en évitant d'utiliser des variants de la littérature qui pourraient introduire un biais dans le résultat de performance comme cela a été observé pour le logiciel Condel dans l'étude de performance dans la thématique DI. De telles études pourraient être réalisées au sein des réseaux de groupes d'expert de pathologie génétique. Cependant, nous ne pouvons écarter que les performances varient également pour les différents gènes d'une même thématique. Il serait alors intéressant de réaliser cette étude sur une cohorte de variants d'un même gène non issus de la littérature. Il serait également judicieux de comparer la performance de ces outils dans une même thématique par rapport à des données issus d'un autre centre afin d'exclure tout biais inter-centre.

Les outils de prédiction de l'effet des mutations faux sens fournissent des informations pouvant aider à la classification des variants, mais ils ne permettent pas de les classer avec certitude car ils ne restent que des outils de prédictions. Les résultats des prédictions doivent être interprétés conjointement avec les autres données clinico- biologiques qui ont une plus grande importance dans la décision finale de classification. Pour confirmer l'implication ou non du variant dans la pathologie, le recours à un test fonctionnel peut

permettre d'apporter un argument supplémentaire. Cependant ces tests sont difficilement accessibles au laboratoire de diagnostic et nécessite pour le moment une étroite collaboration avec les équipes de recherche.

Une autre approche globale pourrait permettre de faire le lien entre les données issues du génome, le phénotype d'un individu et la physiopathologie des maladies. Cette approche combine les différentes stratégies « omiques », la génomique, la transcriptomique, la protéomique et la métabolomique. La combinaison de ces différentes analyses est appelée « étude trans-omique » (Figure 52) (95). L'objectif de ce type d'analyse est de reconstruire des réseaux biochimiques globaux à travers les différentes couches omiques. Cela permet de combiner des technologies omiques dites « statiques » tel que le génome à des approches plus dynamiques tel que l'étude du métabolome (96). Ces études qui relient phénotypes et réseaux trans-omiques permettent de prendre en compte à la fois les facteurs génétiques et environnementaux, ce qui se révèle très utile dans des maladies complexes tel que le diabète de type 2 (97). La confrontation des données génomiques aux autres données « omique » pourrait permettre de suivre l'évolution d'une pathologie afin de mieux en comprendre sa physiopathologie.

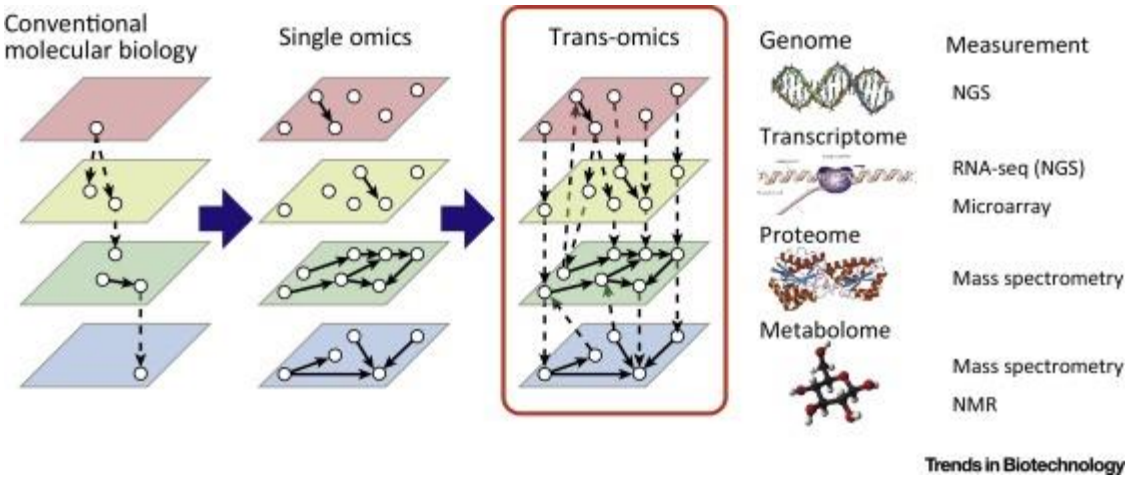

Figure 52. Stratégies « omiques » et intégration au système « trans-omique » (95)**.**

L'essor du séquençage à haut débit nécessite que le biologiste utilise d'autres outils que la génomique. Le développement des techniques de génomique nécessite également le développement d'autres approches « omiques » tel que la protéomique ou la métabolomique afin de les confronter et de proposer la médecine personnalisée la plus exhaustive possible.

V. Références bibliographiques

1. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977 Feb;74(2):560.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463.
3. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977 Feb 24;265(5596):687–95.
4. Ezratty O. Les technologies de séquençage du génome humain – 6 [Internet]. Opinions Libres - Le blog d'Olivier Ezratty. [cited 2017 Jul 6]. Available from: http://www.oezratty.net/wordpress/2012/technologies-sequencage-genome-humain-6/
5. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. Nature. 1986 Jun 12;321(6071):674–9.
6. Dulbecco R. A turning point in cancer research: sequencing the human genome. Science. 1986 Mar 7;231(4742):1055–6.
7. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953 Apr 25;171(4356):737–8.
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.
9. Le Paslier D, Bernot A. Le Projet Génome Humain: quinze ans d'efforts. 2001 [cited 2017 Jul 6]; Available from: http://www.ipubli.inserm.fr/handle/10608/1915
10. Church GM. Genomes for all. Sci Am-Am Ed-. 2006;294(1):46.
11. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011 Feb 10;470(7333):198–203.
12. The Cost of Sequencing a Human Genome [Internet]. National Human Genome Research Institute (NHGRI). [cited 2017 Jul 18]. Available from: https://www.genome.gov/27565109/The-Cost-of-Sequencing-a-Human-Genome
13. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010 Jan;11(1):31–46.
14. lexnederbragt. Developments in high throughput sequencing – July 2016 edition [Internet]. In between lines of code. 2016 [cited 2017 Jul 18]. Available from: https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/
15. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. Genet Med. 2015 Jun;17(6):444–51.

16. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med Off J Am Coll Med Genet. 2013 Jul;15(7):565–74.

17. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015 Apr 4;385(9975):1305–14.

18. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014 Jul 17;511(7509):344–7.

19. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, et al. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. Hum Mutat. 2015 Sep;36(9):903–14.

20. Ion AmpliSeq™ Library Kit 2.0 Workflow [Internet]. [cited 2017 Jul 19]. Available from: https://www.thermofisher.com/fr/fr/home/technical-resources/research-tools/image-gallery/image-gallery-detail.19800.html

21. Samorodnitsky E, Datta J, Jewell BM, Hagopian R, Miya J, Wing MR, et al. Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing. J Mol Diagn. 2015 Jan 1;17(1):64–75.

22. Mardis ER. Next-Generation Sequencing Platforms. Annu Rev Anal Chem. 2013;6(1):287–303.

23. Hager G, Wellein G. Introduction to High Performance Computing for Scientists and Engineers. CRC Press; 2010. 350 p.

24. Oliver GR, Hart SN, Klee EW. Bioinformatics for Clinical Next Generation Sequencing. Clin Chem. 2015 Jan 1;61(1):124–35.

25. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 Apr;38(6):1767.

26. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 2008 May;18(5):763–70.

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

29. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178.

30. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol. 2011 Jan;29(1):24.

31. DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.

32. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

33. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med Off J Am Coll Med Genet. 2015 May;17(5):405–24.

34. Lupski JR, Reid JG, Gonzaga-Jauregui C, Deiros DR, Chen DCY, Nazareth L, et al. Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. N Engl J Med. 2010 Apr 1;362(13):1181–91.

35. Robasky K, Lewis NE, Church GM. The Role of Replicates for Error Mitigation in Next-Generation Sequencing. Nat Rev Genet. 2014 Jan;15(1):56–62.

36. Song W, Gardner SA, Hovhannisyan H, Natalizio A, Weymouth KS, Chen W, et al. Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. Genet Med. 2016 Aug;18(8):850–4.

37. Grandval P, Fabre AJ, Olschwang S. Design of a Core Classification Process for DNA Mismatch Repair Variations of A Priori Unknown Functional Significance. Hum Mutat. 2013 Jun 1;34(6):920–2.

38. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D862–8.

39. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D789.

40. Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat. 2008 Nov;29(11):1327.

41. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812.

42. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. Science. 1974 Sep 6;185(4154):862–4.

43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

44. Luu T-D, Rusu A, Walter V, Linard B, Poidevin L, Ripp R, et al. KD4v: comprehensible knowledge discovery system for missense variant. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W71–5.

45. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915–9.

46. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. Am J Hum Genet. 2011 Apr 8;88(4):440–9.

47. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–5.

48. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug;7(8):575–6.

49. Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, Babaei M, et al. Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: Maximum entropy estimates of splice junction strengths. Hum Mutat. 2004 Jan 1;23(1):67–76.

50. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009 May;37(9):e67.

51. Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, et al. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. Hum Mutat. 2012 Aug 1;33(8):1228–38.

52. Redin C, Gérard B, Lauer J, Herenger Y, Muller J, Quartier A, et al. Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J Med Genet. 2014 Nov;51(11):724–36.

53. Smigiel R, Kostrzewa G, Kosinska J, Pollak A, Stawinski P, Szmida E, et al. Further evidence for GRIN2B mutation as the cause of severe epileptic encephalopathy. Am J Med Genet A. 2016 Dec;170(12):3265–70.

54. Endele S, Rosenberger G, Geider K, Popp B, Tamer C, Stefanova I, et al. Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. Nat Genet. 2010 Nov;42(11):1021–6.

55. Lemke JR, Hendrickx R, Geider K, Laube B, Schwake M, Harvey RJ, et al. GRIN2B Mutations in West Syndrome and Intellectual Disability with Focal Epilepsy. Ann Neurol. 2014 Jan;75(1):147–54.

56. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging. 1989;29(3):307–35.

57. Colombet I, Touzé E. Indices de performance diagnostique. Sang Thromb Vaiss. 2011;23(6):307–316.

58. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerg Themes Epidemiol [Internet]. 2017 Sep 20 [cited 2018 Jan 13];14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5607590/

59. Corcos L, Solier S. Épissage alternatif, pathologie et thérapeutique moléculaire. médecine/sciences. 21(3):253–60.

60. Gaildrat P, Killian A, Martins A, Tournier I, Frébourg T, Tosi M. Use of Splicing Reporter Minigene Assay to Evaluate the Effect on Splicing of Unclassified Genetic Variants. In: Cancer Susceptibility [Internet]. Humana Press, Totowa, NJ; 2010 [cited 2017 Dec 22]. p. 249–57. (Methods in Molecular Biology). Available from: https://link.springer.com/protocol/10.1007/978-1-60761-759-4_15

61. Cooper TA. Use of minigene systems to dissect alternative splicing elements. Methods. 2005 Dec 1;37(4):331–40.

62. Giribaldi G, Doria-Lamba L, Biancheri R, Severino M, Rossi A, Santorelli FM, et al. Intermittent-relapsing pyruvate dehydrogenase complex deficiency: a case with clinical, biochemical, and neuroradiological reversibility. Dev Med Child Neurol. 2012 May 1;54(5):472–6.

63. Jang M-A, Kim BC, Ki C-S, Lee S-Y, Kim J-W, Choi TY, et al. Identification of PRODH Mutations in Korean Neonates with Type I Hyperprolinemia. Ann Clin Lab Sci. 2013 Dec 21;43(1):31–6.

64. Balduyck M, Odou M-F, Zerimech F, Porchet N, Lafitte J-J, Maitre B. Diagnosis of alpha-1 antitrypsin deficiency: Modalities, indications and diagnosis strategy. Rev Mal Respir. 2014 Oct;31(8):729–45.

65. Prins J, Meijden BB van der, Kraaijenhagen RJ, Wielders JPM. Inherited Chronic Obstructive Pulmonary Disease: New Selective-Sequencing Workup for α1-Antitrypsin Deficiency Identifies 2 Previously Unidentified Null Alleles. Clin Chem. 2008 Jan 1;54(1):101–7.

66. Jourdy Y, Chatron N, Fretigny M, Carage ML, Chambost H, Claeyssens-Donadel S, et al. Molecular cytogenetic characterization of five F8 complex rearrangements: utility for haemophilia A genetic counselling. Haemophilia. 2017 Jul 1;23(4):e316–23.

67. Jourdy Y, Nougier C, Roualdes O, Fretigny M, Durand B, Negrier C, et al. Characterization of five associations of F8 missense mutations containing FVIII B domain mutations. Haemophilia. 2016 Jul 1;22(4):583–9.

68. Lassalle F. Mise au point du séquençage haut débit des gènes F8 et F9 dans l'hémophilie A et B par Next-Generation Sequencing. Application aux hémophiles de la région Nord-Pas de Calais. Lille 2;

69. Lynch H, Chapelle A de la. Genetic susceptibility to non-polyposis colorectal cancer. J Med Genet. 1999 Nov;36(11):801.

70. Vasen HFA, Blanco I, Aktan-Collan K, Gopie JP, Alonso A, Aretz S, et al. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. Gut. 2013 Jun;62(6):812.

71. Kazmierczak-Vermaut C. Mise en place d'un test fonctionnel pour évaluer les variants de signification inconnue des gènes MLH1 et MSH2 responsables du syndrome de Lynch [Thèse d'exercice]. [Lille, France]: Université du droit et de la santé; 2014.

72. Strom SP, Lozano R, Lee H, Dorrani N, Mann J, O'Lague PF, et al. De Novo variants in the KMT2A (MLL) gene causing atypical Wiedemann-Steiner syndrome in two unrelated individuals identified by clinical exome sequencing. BMC Med Genet. 2014 May 1;15:49.

73. Agger K, Cloos PAC, Christensen J, Pasini D, Rose S, Rappsilber J, et al. UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature. 2007 Oct 11;449(7163):731–4.

74. Laarhoven PMV, Neitzel LR, Quintana AM, Geiger EA, Zackai EH, Clouthier DE, et al. Kabuki syndrome genes KMT2D and KDM6A: functional analyses demonstrate

critical roles in craniofacial, heart and brain development. Hum Mol Genet. 2015 Aug 1;24(15):4443.

75. Banka S, Lederer D, Benoit V, Jenkins E, Howard E, Bunstone S, et al. Novel KDM6A (UTX) mutations and a clinical and molecular review of the X-linked Kabuki syndrome (KS2). Clin Genet. 2015 Mar 1;87(3):252–8.

76. Tahiliani M, Mei P, Fang R, Leonor T, Rutenberg M, Shimizu F, et al. The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. Nature. 2007 May 31;447(7144):601–5.

77. Vallianatos CN, Iwase S. Disrupted intricacy of histone H3K4 methylation in neurodevelopmental disorders. Epigenomics. 2015;7(3):503–19.

78. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014 Nov 28;346(6213):1258096.

79. Hsu PD, Lander ES, Zhang F. Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell. 2014 Jun 5;157(6):1262–78.

80. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013 Sep;31(9):822–6.

81. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015 May;16(5):299–311.

82. Ye Z-F, Liu X-L, Han Q, Liao H, Dong X-T, Zhu G-H, et al. Functional characterization of PBP1 gene in Helicoverpa armigera (Lepidoptera: Noctuidae) by using the CRISPR/Cas9 system. Sci Rep [Internet]. 2017 Aug 16 [cited 2017 Dec 2];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5559583/

83. Zhu QM, Ko KA, Ture S, Mastrangelo MA, Chen M-H, Johnson AD, et al. Novel Thrombotic Function of a Human SNP in STXBP5 Revealed by CRISPR/Cas9 Gene Editing in MiceHighlights. Arterioscler Thromb Vasc Biol. 2017 Feb 1;37(2):264–70.

84. Stem Cell Quick Reference [Internet]. [cited 2017 Dec 23]. Available from: http://learn.genetics.utah.edu/content/stemcells/quickref/

85. Sangermano R, Bax NM, Bauwens M, van den Born LI, De Baere E, Garanto A, et al. Photoreceptor Progenitor mRNA Analysis Reveals Exon Skipping Resulting from the ABCA4 c.5461-10T→C Mutation in Stargardt Disease. Ophthalmology. 2016 Jun 1;123(6):1375–85.

86. Braun TA, Mullins RF, Wagner AH, Andorf JL, Johnston RM, Bakall BB, et al. Non-exomic and synonymous variants in ABCA4 are an important cause of Stargardt disease. Hum Mol Genet. 2013 Dec 20;22(25):5136–45.

87. Torriano S, Erkilic N, Faugère V, Damodar K, Hamel CP, Roux A-F, et al. Pathogenicity of a novel missense variant associated with choroideremia and its impact on gene replacement therapy. Hum Mol Genet. 2017 Sep 15;26(18):3573–84.

88. Russo FB, Cugola FR, Fernandes IR, Pignatari GC, Beltrão-Braga PCB. Induced pluripotent stem cells for modeling neurological disorders. World J Transplant. 2015 Dec 24;5(4):209–21.

89. Duman D, Tekin M. Autosomal recessive nonsyndromic deafness genes: a review. Front Biosci J Virtual Libr. 2012 Jun 1;17:2213–36.

90. Smith RJ, Ranum PT. Nonsyndromic Hearing Loss and Deafness, DFNA3. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Stephens K, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2018 Jan 8]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1536/

91. Smith RJ, Jones M-KN. Nonsyndromic Hearing Loss and Deafness, DFNB1. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Stephens K, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2018 Jan 8]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1272/

92. Deafness Variation Database [Internet]. [cited 2018 Jan 8]. Available from: http://deafnessvariationdatabase.org/

93. Hereditary Hearing Loss - Hereditary Hearing loss Homepage [Internet]. [cited 2018 Jan 8]. Available from: http://hereditaryhearingloss.org/

94. Hong K, Hu J, Yu J, Brugada R. Concomitant Brugada-like and short QT electrocardiogram linked to SCN5A mutation. Eur J Hum Genet. 2012 Nov;20(11):1189–92.

95. Yugi K, Kubota H, Hatano A, Kuroda S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple ‘Omic’ Layers. Trends Biotechnol. 2016 Apr;34(4):276–90.

96. Bujak R, Struck-Lewicka W, Markuszewski MJ, Kaliszan R. Metabolomics for laboratory diagnostics. J Pharm Biomed Anal. 2015 Sep 10;113:108–20.

97. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012 Mar 16;148(6):1293.

VI. Annexes

Annexe 1 Critères et méthodes de classification des variants selon l'ACMG

Evidence of pathogenicity	Category
Very strong	PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease Caveats: • Beware of genes where LOF is not a known disease mechanism (e.g., *GFAP*, *MYH7*) • Use caution interpreting LOF variants at the extreme 3′ end of a gene • Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact • Use caution in the presence of multiple transcripts
Strong	PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change Example: Val→Leu caused by either G>C or G>T in the same codon Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level
	PS2 De novo (both maternity and paternity confirmed) in a patient with the disease and no family history Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, and so on, can contribute to nonmaternity.
	PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well established.
	PS4 The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls Note 1: Relative risk or OR, as obtained from case–control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance. Note 2: In instances of very rare variants where case–control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.
Moderate	PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation
	PM2 Absent from controls (or at extremely low frequency if recessive) (Table 6) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium Caveat: Population data for insertions/deletions may be poorly called by next-generation sequencing.
	PM3 For recessive disorders, detected in *trans* with a pathogenic variant Note: This requires testing of parents (or offspring) to determine phase.
	PM4 Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants
	PM5 Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before Example: Arg156His is pathogenic; now you observe Arg156Cys Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level.
	PM6 Assumed de novo, but without confirmation of paternity and maternity
Supporting	PP1 Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease Note: May be used as stronger evidence with increasing segregation data
	PP2 Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease
	PP3 Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.) Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of a variant.
	PP4 Patient's phenotype or family history is highly specific for a disease with a single genetic etiology
	PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation

LOF, loss of function; OR, odds ratio.

Evidence of benign impact	Category
Stand-alone	BA1 Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
Strong	BS1 Allele frequency is greater than expected for disorder (see **Table 6**)
	BS2 Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age
	BS3 Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing
	BS4 Lack of segregation in affected members of a family
	Caveat: The presence of phenocopies for common phenotypes (i.e., cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation.
Supporting	BP1 Missense variant in a gene for which primarily truncating variants are known to cause disease
	BP2 Observed in *trans* with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in *cis* with a pathogenic variant in any inheritance pattern
	BP3 In-frame deletions/insertions in a repetitive region without a known function
	BP4 Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
	Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm cannot be counted as an independent criterion. BP4 can be used only once in any evaluation of a variant.
	BP5 Variant found in a case with an alternate molecular basis for disease
	BP6 Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
	BP7 A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

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

Annexe 2 Script R pour analyse statistique complète

Guillaume_GRZYCH

##Importation et définition

```
setwd("~/R/These")
data<-na.omit(read.csv2("Import.csv",dec="."))
library("rpart", lib.loc="~/R/win-library/3.4")
library("rpart.plot")
data$PolyPhen<-as.numeric(as.character(data$PolyPhen))
data$SIFT<-as.numeric(as.character(data$SIFT))
data$MutationTaster_converted_rankscore<-as.numeric(as.character(data$Mut
ationTaster_converted_rankscore))
data$CADD_raw_rankscore<-as.numeric(as.character(data$CADD_raw_rankscore)
)
data$GERP.._RS_rankscore<-as.numeric(as.character(data$GERP.._RS_rankscor
e))
data$PROVEAN_converted_rankscore<-as.numeric(as.character(data$PROVEAN_co
nverted_rankscore))
data$phastCons100way_vertebrate_rankscore<-as.numeric(as.character(data$p
hastCons100way_vertebrate_rankscore))
data$BLOSUM62<-as.numeric(as.character(data$BLOSUM62))
data$Condel<-as.numeric(as.character(data$Condel))
logiciel<-c("SIFT","PolyPhen","Condel","MutationTaster_converted_rankscor
e","PROVEAN_converted_rankscore","BLOSUM62","GERP.._RS_rankscore","phastC
ons100way_vertebrate_rankscore","CADD_raw_rankscore")
```

##Evaluation des performances de chaque outil

```
library(ROCR)

library(pROC)

attach(data)
#Condel
pred=prediction(Condel,Classification)
perf=performance(pred,"tpr", "fpr")
ROCCondel<-plot(perf,colorize = TRUE,main="Condel")

AUC=performance(pred,"auc")
AUCCondel<-AUC@y.values[[1]]
AUCCondel

attr(AUC,"y.values")[[1]]

vx<-perf@x.values[[1]]
vy<-perf@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va<-perf@alpha.values[[1]]
seuilideal=va[which.min(dist)]
```

```
condelSeuil<-(cut(data$Condel,c(0,seuilideal,Inf),labels=c("<seuil",">seu
il")))


tc<-table(data$Classification,condelSeuil)
SeCondel<-(tc[2,2]/(tc[2,2]+tc[2,1])*100)
SpeCondel<-(tc[1,1]/(tc[1,1]+tc[1,2])*100)
YiCONDEL=SeCondel+SpeCondel-100
VPPCondel<-(tc[2,2]/(tc[1,2]+tc[2,2])*100)
VPNCondel<-(tc[1,1]/(tc[1,1]+tc[2,1])*100)
VPPCondel

#PolyPhen
pred2=prediction(PolyPhen,Classification)
perf2=performance(pred2,"tpr", "fpr")
ROCPolyPhen<-plot(perf2,colorize = TRUE,main="PolyPhen")

AUC2=performance(pred2,"auc")
AUCPolyPhen<-AUC2@y.values[[1]]
attr(AUC2,"y.values")[[1]]

vx<-perf2@x.values[[1]]
vy<-perf2@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)

va2<-perf2@alpha.values[[1]]
seuilideal2=va2[which.min(dist)]
PolyPhenSeuil<-(cut(data$PolyPhen,c(0,seuilideal2,Inf),labels=c("<seuil",
">seuil")))
tp<-table(data$Classification,PolyPhenSeuil)
SePolyPhen<-(tp[2,2]/(tp[2,1]+tp[2,2])*100)
SpePolyPhen<-(tp[1,1]/(tp[1,1]+tp[1,2])*100)
YiPolyPhen=SePolyPhen+SpePolyPhen-100
VPPPolyPhen<-(tp[2,2]/(tp[1,2]+tp[2,2])*100)
VPNPolyPhen<-(tp[1,1]/(tp[1,1]+tp[2,1])*100)


#GERP
pred3=prediction(GERP.._RS_rankscore,Classification)
perf3=performance(pred3,"tpr", "fpr")
ROCGERP<-plot(perf3,colorize = TRUE,main="GERP")

AUC3=performance(pred3,"auc")
AUCGERP<-AUC3@y.values[[1]]
attr(AUC3,"y.values")[[1]]

vx<-perf3@x.values[[1]]
vy<-perf3@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va3<-perf3@alpha.values[[1]]
perf3@y.values[[1]]
```

```
seuilideal3=va3[which.min(dist)]
GERPSeuil<-(cut(data$GERP.._RS_rankscore,c(0,seuilideal3,Inf),labels=c("<
seuil",">seuil")))
tg<-table(data$Classification,GERPSeuil)
SeGERP<-(tg[2,2]/(tg[2,1]+tg[2,2])*100)
SpeGERP<-(tg[1,1]/(tg[1,1]+tg[1,2])*100)
YiGERP<-SeGERP+SpeGERP-100


VPPGERP<-(tg[2,2]/(tg[1,2]+tg[2,2])*100)
VPNGERP<-(tg[1,1]/(tg[1,1]+tg[2,1])*100)


#SIFT
pred4=prediction(data$SIFT,Classification)
perf4=performance(pred4,"tpr", "fpr")
ROCSIFT<-plot(perf4,colorize = TRUE,main="SIFT")

AUC4=performance(pred4,"auc")
AUCSIFT<-(auc(roc(Classification,SIFT)))
attr(AUC4,"y.values")[[1]]

vx<-perf4@x.values[[1]]
vy<-perf4@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va4<-perf4@alpha.values[[1]]
seuilideal4=va4[which.min(dist)]
seuilideal4

SIFTseuil<-(cut(data$SIFT,c(0,seuilideal4,Inf),labels=c("<seuil",">seuil"
)))
ts<-table(data$Classification,SIFTseuil)
SeSIFT<-(ts[2,1]/(ts[2,1]+ts[2,2])*100)
SpeSIFT<-(ts[1,2]/(ts[1,1]+ts[1,2])*100)
YiSIFT<-SeSIFT+SpeSIFT-100
VPPSIFT<-(ts[2,1]/(ts[2,1]+ts[1,1])*100)
VPNSIFT<-(ts[1,2]/(ts[1,2]+ts[2,2])*100)


#BLOSUM62
pred5=prediction(BLOSUM62,Classification)
perf5=performance(pred5,"tpr", "fpr")
ROCBLOSUM62<-plot(perf5,colorize = TRUE,main="BLOSUM62")

AUCBLOSUM<-auc(roc(Classification,BLOSUM62))
attr(AUC5,"y.values")[[1]]

vx<-perf5@x.values[[1]]
vy<-perf5@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va5<-perf5@alpha.values[[1]]
```

```
seuilideal5=va5[which.min(dist)]
BLOSUMseuil<-(cut(data$BLOSUM62,c(-4,seuilideal5,Inf),labels=c("<seuil","
>seuil")))
tB<-table(data$Classification,BLOSUMseuil)
SeBLOSUM<-(tB[2,1]/(tB[2,1]+tB[2,2])*100)
SpeBLOSUM<-(tB[1,2]/(tB[1,1]+tB[1,2])*100)
YiBLOSUM<-SeBLOSUM+SpeBLOSUM-100
VPPBLOSUM<-(tB[2,1]/(tB[2,1]+tB[1,1])*100)
VPNBLOSUM<-(tB[1,2]/(tB[1,2]+tB[2,2])*100)

#MutationTaster
pred6=prediction(MutationTaster_converted_rankscore,Classification)
perf6=performance(pred6,"tpr", "fpr")
ROCMutationTaster<-plot(perf6,colorize = TRUE,main="MutationTaster")

AUC6=performance(pred6,"auc")
AUCMutationTaster<-AUC6@y.values[[1]]
attr(AUC6,"y.values")[[1]]

vx<-perf6@x.values[[1]]
vy<-perf6@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va6<-perf6@alpha.values[[1]]
seuilideal6=va6[which.min(dist)]

MutationTasterSeuil<-(cut(data$MutationTaster_converted_rankscore,c(0,seu
ilideal6-0.01,Inf),labels=c("<seuil",">seuil")))
tm<-table(data$Classification,MutationTasterSeuil)
SeMT<-(tm[2,2]/(tm[2,2]+tm[2,1])*100)
SpeMT<-(tm[1,1]/(tm[1,1]+tm[1,2])*100)
YiMutationTaster<-SeMT+SpeMT-100
YiMutationTaster

VPPMT<-(tm[2,2]/(tm[2,2]+tm[1,2])*100)
VPNMT<-(tm[1,1]/(tm[1,1]+tm[2,1])*100)

#PROVEAN
pred7=prediction(PROVEAN_converted_rankscore,Classification)
perf7=performance(pred7,"tpr", "fpr")
ROCProvean<-plot(perf7,colorize = TRUE,main="PROVEAN")

AUC7=performance(pred7,"auc")
AUCProvean<-AUC7@y.values[[1]]
attr(AUC7,"y.values")[[1]]

vx<-perf7@x.values[[1]]
vy<-perf7@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va7<-perf7@alpha.values[[1]]
seuilideal7=va7[which.min(dist)]
```

```
PROVEANSeuil<-(cut(data$PROVEAN_converted_rankscore,c(0,seuilideal7,Inf),
labels=c("<seuil",">seuil")))
tpro<-table(data$Classification,PROVEANSeuil)
SePRO<-(tpro[2,2]/(tpro[2,1]+tpro[2,2])*100)
SpePRO<-(tpro[1,1]/(tpro[1,1]+tpro[1,2])*100)
VPPPRO<-(tpro[2,2]/(tpro[2,2]+tpro[1,2])*100)
VPNPRO<-(tpro[1,1]/(tpro[1,1]+tpro[2,1])*100)


#PhastCons
pred8=prediction(phastCons100way_vertebrate_rankscore,Classification)
perf8=performance(pred8,"tpr", "fpr")
ROCPhas<-plot(perf8,colorize = TRUE,main="PhastCons")

AUC8=performance(pred8,"auc")
AUCPhastCons<-AUC8@y.values[[1]]
attr(AUC8,"y.values")[[1]]

vx<-perf8@x.values[[1]]
vy<-perf8@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va8<-perf8@alpha.values[[1]]
seuilideal8=va8[which.min(dist)]
phastSeuil<-(cut(data$phastCons100way_vertebrate_rankscore,c(0,0.7,Inf),l
abels=c("<seuil",">seuil")))
tph<-table<-table(data$Classification,phastSeuil)
SePhas<-(tph[2,2]/(tph[2,1]+tph[2,2])*100)
SpePhas<-(tph[1,1]/(tph[1,1]+tph[1,2])*100)
SePhas

VPPPhas<-(tph[2,2]/(tph[2,2]+tph[1,2])*100)
VPNPhas<-(tph[1,1]/(tph[1,1]+tph[2,1])*100)
VPPPhas

#CADD
pred9=prediction(CADD_raw_rankscore,Classification)
perf9=performance(pred9,"tpr", "fpr")
ROCCADD<-plot(perf9,colorize = TRUE,main="CADD")

AUC9=performance(pred9,"auc")
AUCCADD<-AUC9@y.values[[1]]
attr(AUC9,"y.values")[[1]]

vx<-perf9@x.values[[1]]
vy<-perf9@y.values[[1]]
dist=sqrt(vx^2+(1-vy)^2)
va9<-perf9@alpha.values[[1]]
seuilideal9=va9[which.min(dist)]
CADDseuil<-(cut(data$CADD_raw_rankscore,c(0,seuilideal9,Inf),labels=c("<s
euil",">seuil")))
tcadd<-table<-table(data$Classification,CADDseuil)
```

```
SeCADD<-(tcadd[2,2]/(tcadd[2,1]+tcadd[2,2])*100)
SpeCADD<-(tcadd[1,1]/(tcadd[1,1]+tcadd[1,2])*100)
VPPCADD<-(tcadd[2,2]/(tcadd[2,2]+tcadd[1,2])*100)
VPNCADD<-(tcadd[1,1]/(tcadd[1,1]+tcadd[2,1])*100)
```

##Représentations graphiques

```
#Afficher toutes les courbes ROC
{par(mfrow = c(3,3))
plot(perf9,colorize = TRUE,main="CADD",ylab="Se",xlab="1-Spe")
plot(perf8,colorize = TRUE,main="PhastCons",ylab="Se",xlab="1-Spe")
plot(perf7,colorize = TRUE,main="PROVEAN",ylab="Se",xlab="1-Spe")
plot(perf6,colorize = TRUE,main="MutationTaster",ylab="Se",xlab="1-Spe")
plot(perf5,colorize = TRUE,main="BLOSUM62",ylab="Se",xlab="1-Spe")
plot(perf4,colorize = TRUE,main="SIFT",ylab="Se",xlab="1-Spe")
plot(perf3,colorize = TRUE,main="GERP",ylab="Se",xlab="1-Spe")
plot(perf2,colorize = TRUE,main="PolyPhen",ylab="Se",xlab="1-Spe")
plot(perf,colorize = TRUE,main="Condel",ylab="Se",xlab="1-Spe")}

par(mfrow = c(1,1))

#Visualisation des performances (Se,Spe,VPP,VPN)
Performance<-data.frame(logiciel = c("Condel","SIFT","PolyPhen","GERP","B
LOSUM","PROVEAN","MutationTaster","PhasCons","CADD"),Se = c(SeCondel,SeSI
FT,SePolyPhen,SeGERP,SeBLOSUM,SePRO,SeMT,SePhas,SeCADD), Spe = c(SpeConde
l,SpeSIFT,SpePolyPhen,SpeGERP,SpeBLOSUM,SpePRO,SpeMT,SpePhas,SpeCADD))


df2 <- data.frame(Performance=rep(c("Se","Spe","VPP","VPN"), each=9),
                  logiciel=rep(c("Condel", "SIFT", "PolyPhen","GERP","BLOSU
M62","PROVEAN","MutationTaster","PhastCons","CADD"),2),
                  Valeur=c(SeCondel, SeSIFT, SePolyPhen, SeGERP, SeBLOSUM,
SePRO,SeMT,SePhas,SpeCADD,SpeCondel, SpeSIFT,SpePolyPhen,SpeGERP,SpeBLOSU
M,SpePRO,SpeMT,SpePhas,SpeCADD,VPPCondel,VPPSIFT,VPPPolyPhen,VPPGERP,VPPB
LOSUM,VPPPRO,VPPMT,VPPPhas,VPPCADD,VPNCondel,VPNSIFT,VPNPolyPhen,VPNGERP,
VPNBLOSUM,VPNPRO,VPNMT,VPNPhas,VPNCADD))
p3<-ggplot(data=df2, aes(x=logiciel, y=Valeur, fill=Performance)) +
geom_bar(stat="identity", position=position_dodge())+coord_flip()
df3 <- data.frame(Performance=rep(c("VPP","VPN"), each=9),
                  logiciel=rep(c("Condel", "SIFT", "PolyPhen","GERP","BLOSU
M62","PROVEAN","MutationTaster","PhastCons","CADD"),2),
                  Valeur=c(VPPCondel,VPPSIFT,VPPPolyPhen,VPPGERP,VPPBLOSUM,
VPPPRO,VPPMT,VPPPhas,VPPCADD,VPNCondel,VPNSIFT,VPNPolyPhen,VPNGERP,VPNBLO
SUM,VPNPRO,VPNMT,VPNPhas,VPNCADD))

p5<-ggplot(data=df3, aes(x=logiciel, y=Valeur, fill=Performance)) +
geom_bar(stat="identity", position=position_dodge())+ggtitle("Performance
des logiciels")+coord_flip()
```

```
df4 <- data.frame(Performance=rep(c("Se","Spe"), each=9),
                  logiciel=rep(c("Condel", "SIFT", "PolyPhen","GERP","BLOSUM62","PROVEAN","MutationTaster","PhastCons","CADD"),2),
                  Valeur=c(SeCondel, SeSIFT, SePolyPhen, SeGERP, SeBLOSUM, SePRO,SeMT,SePhas,SpeCADD,SpeCondel, SpeSIFT,SpePolyPhen,SpeGERP,SpeBLOSUM,SpePRO,SpeMT,SpePhas,SpeCADD))

p6<-ggplot(data=df4, aes(x=logiciel, y=Valeur, fill=Performance)) + geom_bar(stat="identity", position=position_dodge())+ggtitle("Performance des logiciels")
```

##Arbre de décision, classement des outils et utilisation

```
#Arbre de décision, méthode CART
library(rpart)
library(rpart.plot)
attach(data)
tree<-rpart(Classification~.,data=data[logiciel],minsplit=10,control=rpart.control(cp=0,0))
prp(tree,extra=1,main="Arbre de classement des logiciels")
```

Faculté des Sciences Pharmaceutiques et Biologiques de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
☎ 03.20.96.40.40 - Télécopie : 03.20.96.43.64
http://pharmacie.univ-lille2.fr/

Université de Lille 2 DROIT ET SANTÉ

DEMANDE D'AUTORISATION DE SOUTENANCE

Nom et Prénom de l'étudiant : GRZYCH Guillaume

Date, heure et lieu de soutenance :

Le 16/02/2018 (jour / mois / année) à 18h15 Amphithéâtre ou salle : CURIE

Avis du conseiller (directeur) de thèse

Nom : Leclerc Prénom : Julie

☒ Favorable

☐ Défavorable

Motif de l'avis défavorable :

Date : 12/01/18

Signature:

Avis du Président de Jury

Nom : Brousseau Prénom : Thierry

☒ Favorable

☐ Défavorable

Motif de l'avis défavorable :

Date : 11-01-2018

Signature:

Décision de Monsieur le Doyen

☒ Favorable

☐ Défavorable

Le Doyen

D. CUNY

NB : La faculté n'entend donner aucune approbation ou improbation aux opinions émises dans les thèses, qui doivent être regardées comme propres à leurs auteurs.

NA/ 2015

Faculté des Sciences Pharmaceutiques et Biologiques de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
☎ 03.20.96.40.40 - Télécopie : 03.20.96.43.64
http://pharmacie.univ-lille2.fr/

Université de Lille 2 DROIT ET SANTÉ

DEMANDE D'AUTORISATION DE SOUTENANCE

Nom et Prénom de l'étudiant : GRZYCH Guillaume

Date, heure et lieu de soutenance :

Le |1|6| |0|2| |2|0|1|8| (jour / mois / année) à 18 h 15 Amphithéâtre ou salle : CURIE

Avis du conseiller (directeur) de thèse

Nom : Ghoumid Prénom : Jamal

☒ Favorable

☐ Défavorable

Motif de l'avis défavorable :.....

.....

.....

Date : 12-01-2018
Signature:

Avis du Président de Jury

Nom : Broussseau Prénom : Thierry

☒ Favorable

☐ Défavorable

Motif de l'avis défavorable :.....

.....

.....

Date : 11-01-2018
Signature:

Décision de Monsieur le Doyen

☒ Favorable

☐ Défavorable

Le Doyen

D. CUNY

NB : La faculté n'entend donner aucune approbation ou improbation aux opinions émises dans les thèses, qui doivent être regardées comme propres à leurs auteurs.

NA/ 2015

Université de Lille

FACULTE DE PHARMACIE DE LILLE

MEMOIRE de DIPLOME D'ETUDES SPECIALISEES
(tenant lieu de Thèse en vue du Diplôme d'Etat de Docteur en Pharmacie)
Année Universitaire 2017/2018

Nom : GRZYCH

Prénom : Guillaume

Titre de la thèse :

Evaluation des outils de prédiction in silico et intérêt des tests fonctionnels dans l'interprétation des variants identifiés par séquençage de nouvelle génération en génétique humaine.

Mots-clés : Génétique humaine, Séquençage haut débit, Bioinformatique, Analyses *in silico*, tests fonctionnels.

Résumé : L'utilisation du séquençage de nouvelle génération tend à se généraliser dans les laboratoires de génétique moléculaire. En effet, cette technologie permet l'analyse rapide d'un grand nombre de gènes chez un grand nombre de patients, dans des délais relativement courts. Toutefois, la contrepartie est l'identification d'un très grand nombre de variants par gène étudié et par patient. A l'échelle d'un génome entier, plus de 3 millions de variations sont retrouvées. Cependant la grande majorité n'auront pas d'effet délétère. Il convient donc d'avoir une méthode robuste de caractérisation des variants afin de déterminer lequel ou lesquels sont responsables de la pathologie. Une des étapes clés du traitement bioinformatique des données est « l'annotation » des variants, qui permet de décrire le variant selon plusieurs critères, incluant sa fréquence dans la population générale, sa présence dans les bases de données ou encore les prédictions in silico quant à leur pathogénicité. En prenant en compte tous ces éléments et le phénotype clinique du patient, le biologiste interprète et classe les variants identifiés selon des recommandations telles que celles de l'American College of Medical Genetics and Genomics (ACMG), en cinq classes différentes, allant de « variant pathogène» à «variant bénin»).

Dans certains cas, il n'y a pas suffisamment de preuves pour déterminer si les variants identifiés sont responsables de la pathologie. L'utilisation d'outils de prédiction in silico peut aider le biologiste dans l'interprétation. Cependant ces outils ne suffisent pas à eux seuls et sont parfois peu informatifs. Le variant est alors classé en variant « de signification inconnue », et ce malgré les réunions de concertations pluridisciplinaires (RCP). Afin de sortir de cette impasse, le recours à des analyses fonctionnelles est nécessaire pour mieux comprendre l'impact du variant sur l'expression ou la fonction de la protéine, et donc son implication dans le phénotype du patient. Nous souhaitons faire un état des lieux des pratiques et des stratégies utilisées au sein du CHRU de Lille, d'une part vis-à-vis de l'utilisation des outils in silico dans l'interprétation de ces variants, mais également vis-à-vis des tests fonctionnels afin de savoir dans quelles mesures ceux-ci peuvent être intégrés dans le processus de diagnostic. L'objectif est de déterminer d'une part la performance des outils in silico utilisées et d'autre part de donner quelques exemples de tests fonctionnels qui aident à la classification des variants identifiés en NGS.

Membres du jury :

Président :	Monsieur le Professeur Thierry BROUSSEAU Faculté de Pharmacie, Université de Lille II
Assesseur(s) :	Madame le Professeur Marie-Pierre BUISINE Faculté de Médecine, Université de Lille II
Directeurs de mémoire :	Madame le Docteur Julie LECLERC Faculté de Médecine, Université de Lille II Monsieur le Docteur Jamal GHOUMID Faculté de Médecine, Université de Lille II