Université de Poitiers
Faculté de Médecine et Pharmacie

ANNEE 2019

MEMOIRE
DU DIPLOME D'ETUDES SPECIALISEES
DE BIOLOGIE MEDICALE
(décret du 23 janvier 2003)

et

THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE
(décret du 16 janvier 2004)

présentée et soutenue publiquement
le vendredi 5 Juillet 2019 à Poitiers
par Melle Lucile LETIENNE

Intérêt de la technique de l'exome dans la recherche de nouveaux variants géniques impliqués dans la déficience intellectuelle

Composition du Jury

Président : Madame le Professeur Brigitte Gilbert-Dussardier

Membres : Madame le Professeur Guylène Page
Monsieur le Docteur Frédéric Bilan
Monsieur le Docteur Gwenaël Le Guyader

Directeur de thèse : Monsieur le Docteur Frédéric Bilan

Université de Poitiers
Faculté de Médecine et Pharmacie

ANNEE 2019

MEMOIRE
DU DIPLOME D'ETUDES SPECIALISEES
DE BIOLOGIE MEDICALE
(décret du 23 janvier 2003)

et

THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE
(décret du 16 janvier 2004)

présentée et soutenue publiquement
le vendredi 5 Juillet 2019 à Poitiers
par Melle Lucile LETIENNE

Intérêt de la technique de l'exome dans la recherche de nouveaux variants géniques impliqués dans la déficience intellectuelle

Composition du Jury

Président : Madame le Professeur Brigitte Gilbert-Dussardier

Membres : Madame le Professeur Guylène Page
Monsieur le Docteur Frédéric Bilan
Monsieur le Docteur Gwenaël Le Guyader

Directeur de thèse : Monsieur le Docteur Frédéric Bilan

UNIVERSITE DE POITIERS

Faculté de Médecine et de Pharmacie

Le Doyen,

Année universitaire 2018 - 2019

LISTE DES ENSEIGNANTS DE MEDECINE

Professeurs des Universités-Praticiens Hospitaliers

- ALLAL Joseph, thérapeutique
- BATAILLE Benoît, neurochirurgie **(retraite 09/2019)**
- BRIDOUX Frank, néphrologie
- BURUCOA Christophe, bactériologie – virologie
- CARRETIER Michel, chirurgie générale **(retraite 09/2019)**
- CHEZE-LE REST Catherine, biophysique et médecine nucléaire
- CHRISTIAENS Luc, cardiologie
- CORBI Pierre, chirurgie thoracique et cardio-vasculaire
- DAHYOT-FIZELIER Claire, anesthésiologie – réanimation
- DEBAENE Bertrand, anesthésiologie réanimation
- DEBIAIS Françoise, rhumatologie
- DROUOT Xavier, physiologie
- DUFOUR Xavier, Oto-Rhino-Laryngologie
- FAURE Jean-Pierre, anatomie
- FRASCA Denis, anesthésiologie-réanimation
- FRITEL Xavier, gynécologie-obstétrique
- GAYET Louis-Etienne, chirurgie orthopédique et traumatologique
- GERVAIS Elisabeth, rhumatologie
- GICQUEL Ludovic, pédopsychiatrie
- GILBERT Brigitte, génétique
- GOMBERT Jean-Marc, immunologie
- GOUJON Jean-Michel, anatomie et cytologie pathologiques
- GUILLEVIN Rémy, radiologie et imagerie médicale
- HAUET Thierry, biochimie et biologie moléculaire
- HOUETO Jean-Luc, neurologie
- INGRAND Pierre, biostatistiques, informatique médicale
- JAAFARI Nematollah, psychiatrie d'adultes
- JABER Mohamed, cytologie et histologie
- JAYLE Christophe, chirurgie thoracique et cardio-vasculaire
- KARAYAN-TAPON Lucie, cancérologie
- KEMOUN Gilles, médecine physique et de réadaptation **(en détachement)**
- KRAIMPS Jean-Louis, chirurgie générale
- LECLERE Franck, chirurgie plastique, reconstructrice
- LECRON Jean-Claude, biochimie et biologie moléculaire
- LELEU Xavier, hématologie
- LEVARD Guillaume, chirurgie infantile
- LEVEQUE Nicolas, bactériologie-virologie
- LEVEZIEL Nicolas, ophtalmologie
- MACCHI Laurent, hématologie
- MCHEIK Jiad, chirurgie infantile
- MEURICE Jean-Claude, pneumologie
- MIGEOT Virginie, santé publique
- MILLOT Frédéric, pédiatrie, oncologie pédiatrique
- MIMOZ Olivier, anesthésiologie – réanimation
- NEAU Jean-Philippe, neurologie
- ORIOT Denis, pédiatrie
- PACCALIN Marc, gériatrie
- PERAULT Marie-Christine, pharmacologie clinique
- PERDRISOT Rémy, biophysique et médecine nucléaire
- PIERRE Fabrice, gynécologie et obstétrique
- PRIES Pierre, chirurgie orthopédique et traumatologique
- RICHER Jean-Pierre, anatomie
- RIGOARD Philippe, neurochirurgie
- ROBERT René, réanimation
- ROBLOT France, maladies infectieuses, maladies tropicales
- ROBLOT Pascal, médecine interne
- RODIER Marie-Hélène, parasitologie et mycologie
- SAULNIER Pierre-Jean, thérapeutique
- SCHNEIDER Fabrice, chirurgie vasculaire
- SILVAIN Christine, hépato-gastro- entérologie
- TASU Jean-Pierre, radiologie et imagerie médicale
- THIERRY Antoine, néphrologie
- THILLE Arnaud, réanimation
- TOUGERON David, gastro-entérologie
- TOURANI Jean-Marc, cancérologie **(retraite 09/2019)**
- WAGER Michel, neurochirurgie
- XAVIER Jean, pédopsychiatrie

Maîtres de Conférences des Universités-Praticiens Hospitaliers

- ALBOUY-LLATY Marion, santé publique
- BEBY-DEFAUX Agnès, bactériologie – virologie
- BEN-BRIK Eric, médecine du travail **(en détachement)**
- BILAN Frédéric, génétique
- BOURMEYSTER Nicolas, biologie cellulaire
- CASTEL Olivier, bactériologie - virologie – hygiène
- COUDROY Rémy, réanimation **(en mission 1 an)**
- CREMNITER Julie, bactériologie – virologie
- DIAZ Véronique, physiologie
- FROUIN Eric, anatomie et cytologie pathologiques
- GARCIA Magali, bactériologie-virologie **(en mission 1 an)**
- JAVAUGUE Vincent, néphrologie
- LAFAY Claire, pharmacologie clinique
- PALAZZO Paola, neurologie **(pas avant janvier 2019)**
- PERRAUD Estelle, parasitologie et mycologie
- RAMMAERT-PALTRIE Blandine, maladies infectieuses
- SAPANET Michel, médecine légale
- THUILLIER Raphaël, biochimie et biologie moléculaire

Professeur des universités de médecine générale

- BINDER Philippe
- GOMES DA CUNHA José

Bât. D1 - 6 rue de la Milétrie – TSA 51115 - 86073 POITIERS CEDEX 9 - France
05.49.45.43.43 - 05.49.45.43.05

Professeurs associés de médecine générale

- BIRAULT François
- FRECHE Bernard
- MIGNOT Stéphanie
- PARTHENAY Pascal
- VALETTE Thierry

Maîtres de Conférences associés de médecine générale

- AUDIER Pascal
- ARCHAMBAULT Pierrick
- BRABANT Yann
- VICTOR-CHAPLET Valérie

Enseignants d'Anglais

- DEBAIL Didier, professeur certifié
- GAY Julie, professeur agrégé

Professeurs émérites

- DORE Bertrand, urologie (08/2020)
- EUGENE Michel, physiologie (08/2019)
- GIL Roger, neurologie (08/2020)
- GUILHOT-GAUDEFFROY François, hématologie et transfusion (08/2020)
- HERPIN Daniel, cardiologie (08/2020)
- KITZIS Alain, biologie cellulaire (16/02/2019)
- MARECHAUD Richard, médecine interne (24**/**11/2020**)**
- MAUCO Gérard, biochimie et biologie moléculaire (08/2021)
- RICCO Jean-Baptiste, chirurgie vasculaire (08/2020)
- SENON Jean-Louis, psychiatrie d'adultes (08/2020)
- TOUCHARD Guy, néphrologie (08/2021)

Professeurs et Maîtres de Conférences honoraires

- AGIUS Gérard, bactériologie-virologie
- ALCALAY Michel, rhumatologie
- ARIES Jacques, anesthésiologie-réanimation
- BABIN Michèle, anatomie et cytologie pathologiques
- BABIN Philippe, anatomie et cytologie pathologiques
- BARBIER Jacques, chirurgie générale (ex-émérite)
- BARRIERE Michel, biochimie et biologie moléculaire
- BECQ-GIRAUDON Bertrand, maladies infectieuses, maladies tropicales (ex-émérite)
- BEGON François, biophysique, médecine nucléaire
- BOINOT Catherine, hématologie – transfusion
- BONTOUX Daniel, rhumatologie (ex-émérite)
- BURIN Pierre, histologie
- CASTETS Monique, bactériologie -virologie – hygiène
- CAVELLIER Jean-François, biophysique et médecine nucléaire
- CHANSIGAUD Jean-Pierre, biologie du développement et de la reproduction
- CLARAC Jean-Pierre, chirurgie orthopédique
- DABAN Alain, cancérologie radiothérapie (ex-émérite)
- DAGREGORIO Guy, chirurgie plastique et reconstructrice
- DESMAREST Marie-Cécile, hématologie
- DEMANGE Jean, cardiologie et maladies vasculaires
- FAUCHERE Jean-Louis, bactériologie-virologie (ex-émérite)
- FONTANEL Jean-Pierre, Oto-Rhino Laryngologie (ex-émérite)
- GRIGNON Bernadette, bactériologie
- GUILLARD Olivier, biochimie et biologie moléculaire
- GUILLET Gérard, dermatologie
- JACQUEMIN Jean-Louis, parasitologie et mycologie médicale
- KAMINA Pierre, anatomie (ex-émérite)
- KLOSSEK Jean-Michel, Oto-Rhino-Laryngologie
- LAPIERRE Françoise, neurochirurgie (ex-émérite)
- LARSEN Christian-Jacques, biochimie et biologie moléculaire
- LEVILLAIN Pierre, anatomie et cytologie pathologiques
- MAGNIN Guillaume, gynécologie-obstétrique (ex-émérite)
- MAIN de BOISSIERE Alain, pédiatrie
- MARCELLI Daniel, pédopsychiatrie (ex-émérite)
- MARILLAUD Albert, physiologie
- MENU Paul, chirurgie thoracique et cardio-vasculaire (ex-émérite)
- MORICHAU-BEAUCHANT Michel, hépato-gastro-entérologie
- MORIN Michel, radiologie, imagerie médicale
- PAQUEREAU Joël, physiologie
- POINTREAU Philippe, biochimie
- POURRAT Olivier, médecine interne (ex-émérite)
- REISS Daniel, biochimie
- RIDEAU Yves, anatomie
- SULTAN Yvette, hématologie et transfusion
- TALLINEAU Claude, biochimie et biologie moléculaire
- TANZER Joseph, hématologie et transfusion (ex-émérite)
- VANDERMARCQ Guy, radiologie et imagerie médicale

Bât. D1 - 6 rue de la Milétrie – TSA 51115 - 86073 POITIERS CEDEX 9 - France
☎ 05.49.45.43.43 - 05.49.45.43.05

UNIVERSITE DE POITIERS

Faculté de Médecine et de Pharmacie

Année universitaire 2018-2019

PHARMACIE

Professeurs

- COUET William, pharmacie clinique PU-PH
- DUPUIS Antoine, pharmacie clinique PU-PH
- MARCHAND Sandrine, pharmacocinétique PU-PH
- RAGOT Stéphanie, santé publique PU-PH

- CARATO Pascal, chimie thérapeutique PR
- FAUCONNEAU Bernard, toxicologie PR
- GUILLARD Jérôme, pharmacochimie PR
- IMBERT Christine, parasitologie PR
- OLIVIER Jean Christophe, galénique PR
- PAGE Guylène, biologie cellulaire PR
- RABOUAN Sylvie, chimie physique, chimie analytique PR
- SARROUILHE Denis, physiologie PR
- SEGUIN François, biophysique, biomathématiques PR

Maîtres de Conférences

- BARRA Anne, immunologie-hématologie MCU-PH
- THEVENOT Sarah, hygiène et santé publique MCU-PH

- BARRIER Laurence, biochimie MCF
- BODET Charles, bactériologie MCF
- BON Delphine, biophysique MCF
- BRILLAULT Julien, pharmacocinétique, biopharmacie MCF
- BUYCK Julien, microbiologie, MCF
- CHARVET Caroline, physiologie MCF
- DEBORDE-DELAGE Marie, sciences physico-chimiques MCF
- DELAGE Jacques, biomathématiques, biophysique MCF
- FAVOT-LAFORGE Laure, biologie cellulaire et moléculaire MCF
- GIRARDOT Marion, biologie végétale et pharmacognosie, MCF
- GREGOIRE Nicolas, pharmacologie MCF
- HUSSAIN Didja, pharmacie galénique MCF
- INGRAND Sabrina, toxicologie MCF
- MARIVINGT-MOUNIR Cécile pharmacochimie MCF
- PAIN Stéphanie, toxicologie MCF
- RIOUX BILAN Agnès, biochimie MCF
- TEWES Frédéric, chimie et pharmacochimie MCF
- THOREAU Vincent, biologie cellulaire MCF
- WAHL Anne, chimie analytique MCF

Maîtres de Conférences Associés - officine

- DELOFFRE Clément, pharmacien
- HOUNKANLIN Lydwin, pharmacien

Enseignants d'anglais

- DEBAIL Didier
- GAY Julie

Bât. D1 - 6 rue de la Milétrie – TSA 51115 - 86073 POITIERS CEDEX 9 - France
05.49.45.43.43 - 05.49.45.43.05

REMERCIEMENTS

Aux Membres du Jury :

A la Présidente du Jury, Madame le Professeur Brigitte Gilbert-Dussardier, pour m'avoir permis d'intégrer l'équipe de Génétique Biologique il y a maintenant deux ans et pour quelques années encore je l'espère. C'est pour moi un grand honneur que vous présidiez ce jury. Je vous prie de trouver ici l'expression de mon immense respect.

A Madame le Professeur Guylène Page, pour avoir accepté de faire partie de ce jury et pour m'avoir permis d'effectuer mon stage de Master 2 au sein de l'Unité NEUVACOD. Sans votre aide, je n'aurais pas pu obtenir ce diplôme. Veuillez trouver ici le témoignage de ma profonde reconnaissance.

A Monsieur le Docteur Frédéric Bilan, mon directeur de thèse, pour m'avoir transmis ses connaissances au cours de ces 2 dernières années et pour m'avoir guidé dans ce projet. Il n'aurait pas pu voir le jour sans votre aide. Je vous remercie de m'offrir la chance d'accroître mes connaissances à votre contact et de me permettre d'exercer en tant que biologiste dans le laboratoire de Génétique Biologique. Soyez assuré de ma profonde gratitude.

A Monsieur le Docteur Gwenaël Le Guyader, pour son soutien dans les dernières étapes de cette thèse et m'avoir donné ses conseils avisés concernant le versant clinique de la génétique. Merci à toi pour toutes les explications aussi bien techniques que cliniques, j'espère que notre collaboration pourra encore durer au moins quelques années.

A l'ensemble du Laboratoire de Génétique Biologique :

Aux techniciennes, Barbara, Valérie C, Valérie L, Patricia, Marlène, Betty, Bernadette, Laura et Martine,

Aux ingénieures Montserrat et Sylvie,

Au bio-informaticien (thésard à ses heures perdues) Quentin,

Aux biologistes, Fabienne, Dominique et Matthieu et Frédéric:

Merci à vous tous pour m'avoir accueilli si chaleureusement dans ce laboratoire. Au cours de ma formation, des différents projets auxquels j'ai pu participer, vous avez toujours été là pour me conseiller, pour m'en apprendre plus sur les techniques et le fonctionnement du laboratoire, pour me soutenir ne serait-ce que par un sourire ou une anecdote. Je n'oublie pas que si j'en suis arrivée là aujourd'hui, je le dois aussi à chacun d'entre vous. Je n'ai pas toujours de chocolat sous la main pour vous remercier et les mots sont bien faibles, mais je vous adresse un grand merci à vous tous.

A l'ensemble du Service de Génétique Clinique :

A tous ceux que j'ai croisés au détour des couloirs

Aux conseillères en génétique, Emeline et Aurore,

Aux cliniciens, Brigitte, Gwenaël et Pascaline

Merci pour votre implication auprès des patients, merci pour les collaborations et les échanges étroits que vous entretenez avec le laboratoire. Je vous remercie spécialement pour tous les renseignements que vous m'avez fournis pour ce projet. C'est un grand plaisir de travailler avec vous, votre enthousiasme fait du service de génétique la grande famille qu'elle est aujourd'hui.

Au Service de Cancérologie Biologique

Aux techniciens, Marine, Mélanie, Margaux, Marion, Sébastien et Pierre

Aux ingénieurs, Ulrich et Tristan

A Birama et à Gwenaëlle

Aux biologistes, Gaëlle, Claire et au Pr Lucie Karayan-Tapon

Je vous remercie de m'avoir accueillie dans votre laboratoire il y a deux mois maintenant et de me faire découvrir le côté « somatique » qui manquait à ma formation. Merci de me proposer d'intégrer votre équipe dans les mois à venir, pour le meilleur je l'espère.

A mes amis,

A ma meilleure amie, ma sœur de cœur, Bettina, pour son soutien inconditionnel et toutes les conneries qu'on s'est racontées pendant ces 12 dernières années (Quoi ? Déjà ?). Maintenant, médecine c'est fini, et je suis heureuse que tu aies enfin trouvé le bonheur !

A mes co-internes de génétique, Xavier et Tanguy, eux aussi très « mordus » de génétique (pitié plus de citations de OSS !), et Caroline, qui n'aime « que les chromonomosomes », sans parler de Donovan ou de Fnéda… Merci de m'avoir supporté dans notre petit bureau, malgré des jours pas toujours faciles.

A mes co-internes de bactériologie, Lauranne, Natacha et Valentin, mes amis dans la vie. Les fous-rires, les ragots, les restos, j'espère qu'il y en aura encore assez pour nous occuper ! Merci pour votre soutien.

A mes co-internes de cancérologie biologique, Pierre et Margaux, qui découvrent ou redécouvrent la biologie et qui me rappellent l'importance du patient derrière les analyses.

A mes amis « de l'extérieur », Cha, Mika, Eloïse, Elsa, Gauthier, Adeline, Antoine… et tous ceux qui m'ont accueillie à Poitiers il y a quelques années déjà, qui ont partagé les joies et les peines, et quelques bons coups de « pouët » ! Cette année fut pleine de rebondissements grâce à vous aussi.

A mes parents,

Merci de m'avoir toujours soutenue en me laissant voler de mes propres ailes. Vous avez fait aussi de moi ce que je suis aujourd'hui, j'espère vous rendre fiers.

A mon Amour,

Romain, merci d'être à mes côtés dans tous les jours de cette année bien chargée… Cela n'a pas toujours été simple mais tu as su trouver les mots pour me faire rire, pour me réconforter, pour me booster, pour me féliciter. Tu me donnes la force d'avancer toujours un peu plus loin. Avec ton soutien, si c'était à refaire, je te dirai encore « Oui ! ».

SOMMAIRE

LISTE DES ABREVIATIONS

ACMG : *American College of Medical Genetics*
ACPA : Analyse Chromosomique sur Puce à ADN
AD : Autosomique dominant
ADN : Acide désoxyribonucléique
ADNlcT21 : ADN libre circulant de la trisomie 21
AR : Autosomique récessif
ARN : Acide ribonucléique
.BAM : *Binary Alignment Map*
CGH : *Comparative Genomic Hybridization*
CHARGE : Acronyme anglais pour ***C****oloboma,* ***H****eart defect,* ***A****tresia of the choanae,* ***R****etarded growth and development,* ***G****enital hypoplasia,* ***E****ar anomalies/deafness*
CNV : *Copy Number Variation*
CPU : *Central Processing Unit*
DI : Déficience intellectuelle
DIL : Déficience intellectuelle légère
DIS : Déficience intellectuelle sévère
DPNI : Dépistage Prénatal Non Invasif
Epipage : Etude épidémiologique sur les petits âges gestationnels
ESHG : *European Society of Human Genetics*
ESRP1 : *Epithelial splicing regulatory protein 1*
ERAD : *Endoplasmic reticulum associated degradation*
FASD : *Fetal Alcohol Spectrum Disorder*
FKBP4 : *T-cell FK-506-binding protein*
FMRP : *Fragile X mental retardation protein*
GABA : Acide γ-aminobutyrique
Gb : Gigabases
HAS : Haute Autorité de Santé
HS : Haute Sensibilité
Mb : Mégabases
NCKAP1 : *NCK-associated protein 1*
NGS : *Next Generation Sequencing*
OMIM : *Online Mendelian Inheritance in Man*
OMS : Organisation Mondiale de la Santé
pb : paire de bases
PHRC : Programme Hospitalier de Recherche Clinique
PolyPhen : *Polymorphism Phenotyping*
QI : Quotient intellectuel
RA : Retard des acquisitions
RAM : *Random Access Memory*
REVEL : *Rare Exome Variant Ensemble Learner*
RFID : *Radio Frequency Identification*
RHEOP : Registre des Handicaps de l'Enfant et Observatoire Périnatal de l'Isère et des deux Savoie
RIHN : Référentiel des actes Innovants Hors Nomenclature
rpm : tours par minute
SA : Semaines d'aménorrhée
SAF : Syndrome d'alcoolisation fœtale
SAM : *Sequence Alignment/Map*
SIFT : *Sorting Intolerant From Tolerant*
To : Téraoctet
UPR : *Unfolded protein response*
USP19 : *Ubiquitin-specific protease*
.VCF : *Variant Call Format*
VUS : *Variant of Unknown Significance*

LISTE DES FIGURES

LISTE DES TABLEAUX

I. Introduction

La déficience intellectuelle (DI) est aujourd'hui une véritable question de santé publique. Dans le monde, on estime la prévalence de la DI à 1% tous pays confondus (Maulik *et al.*, 2011). Evaluer non seulement le nombre de personnes avec déficience intellectuelle, mais aussi celui des personnes en situation de handicap, permet aux gouvernements de prendre la mesure des enjeux (sanitaires, éducatifs, sociaux…) afin d'établir une politique de développement des services, de prévoir et de mettre à disposition les ressources nécessaires à la prise en charge de ces personnes.

L'accroissement constant de la population, particulièrement dans les pays en voie de développement, rend nécessaire le développement de moyens de prévention et de dépistage de la DI (Dave *et al.*, 2005; Gustavson, 2005). En effet, un certain nombre de causes de la DI sont évitables, notamment les infections prénatales, ce qui nécessite une prise en charge précoce et des moyens diagnostiques et thérapeutiques adaptés.

L'identification d'un retard développemental chez un enfant a un impact considérable sur son environnement familial, les parents se retrouvant souvent démunis face à cette découverte. La recherche et l'identification d'une étiologie permettant d'expliquer la DI sont importantes pour ces familles, bien que difficiles au vu des facteurs multiples qui peuvent intervenir.

La démarche diagnostique demeure nécessaire puisqu'elle mène à une reconnaissance des difficultés de l'individu (et potentiellement de celles rencontrées par sa famille) et à la mise en place de moyens matériels et humains (orthophonistes, psychomotriciens, psychologues, associations de patients) qui lui permettent de faire face à ses difficultés.

Lors de l'identification d'une cause génétique, le conseil génétique donné lors d'une consultation spécialisée est adapté à la pathologie identifiée ; il doit permettre d'orienter le patient et sa famille vers des solutions afin d'optimiser sa prise en charge. Ceci permet aussi la mise en place d'un suivi médical adapté dans le but de prévenir les complications connues associées à sa pathologie.

Par ailleurs, on estime qu'environ un tiers des 25 000 gènes humains sont exprimés au niveau du cerveau, participant à son développement et à son fonctionnement (Colantuoni *et al.*, 2000). C'est sans doute l'une des explications de l'extrême hétérogénéité génétique des déficiences intellectuelles, l'altération de l'expression ou de la fonction de l'un ou l'autre de ces gènes pouvant affecter le développement cognitif.

A ce jour, plus de 1200 gènes ont été identifiés dans la DI d'après la base de données SysID et de nouveaux gènes sont mis en évidence tous les mois, ce qui rend d'autant plus difficile le diagnostic étiologique.

Dans l'étude des causes génétiques de la DI, la recherche de nouveaux variants ou de nouveaux gènes impliqués dans la DI apparaît comme primordiale. Elle s'appuie désormais sur l'étude de l'exome, c'est-à-dire de la partie du génome qui contient les exons, séquences reconnues codantes pour l'information génétique. Grâce aux techniques NGS, *Next Generation Sequencing*, ce séquençage permet d'accéder pour un coût modeste en termes de réactifs (moins de 600 €) à de très nombreuses variations géniques à l'échelle du nucléotide (substitution, délétion ou insertion d'une ou plusieurs bases nucléotidiques) ou à l'échelle du CNV (C*opy Number Variation*).

Cette innovation technologique couplée aux avancées informatiques génère une véritable « carte génique » à l'échelle moléculaire de l'individu. Par son analyse, il est possible de répertorier tous les variants géniques propres à chaque individu, et de les comparer à des bases de données. Par différentes approches informatiques, il est alors possible de « filtrer » ces variants et de les classer en fonction de leur pathogénicité supposée.

Il s'agit alors d'essayer de corréler la présence des variants classés comme pathogènes au phénotype particulier présenté par l'individu et ainsi, soit de poser un diagnostic si le variant et/ou le gène incriminé a déjà été impliqué dans la DI, soit de mettre en évidence un nouveau gène.

Il est à noter que cet examen n'est toujours pas référencé dans la liste des actes innovants de la biologie médicale (référentiel RIHN : Référentiel des actes Innovants Hors Nomenclature). D'autres panels de gènes ciblés, allant jusqu'à l'analyse NGS de 600 gènes connus dans la DI, sont pourtant référencés dans le RIHN 2019 et codés avec une valorisation se situant entre 880 et 2200 euros environ.

Au vu du nombre de gènes impliqués dans la DI, la réalisation de plusieurs panels chez un même patient n'est pas envisageable, d'où l'intérêt de l'étude de l'exome, qui permet non seulement d'identifier des variations pathogènes dans des gènes bien connus, mais aussi, dans le cadre de la recherche, des variations dans des gènes non encore répertoriés, candidats dans la DI.

Dans ce cadre, nous effectuerons deux types d'études en parallèle. De manière rétrospective, nous analyserons des données d'exomes (dont la réaction de séquençage a déjà été réalisée) provenant de différents projets de recherche, tels les projets HUGODIMS 1 et 2 (Projet Inter-régional du Grand Ouest pour l'exploration par approche exome des causes moléculaires de DI isolée ou syndromique Modérée et Sévère). A partir de ces données, nous tenterons de définir la meilleure stratégie d'analyse informatique afin d'identifier de nouveaux variants géniques.

De plus, de manière prospective, nous effectuerons un séquençage de l'exome pour une cinquantaine de patients issus des consultations de génétique du CHU de Poitiers et nous les analyserons selon le protocole informatique établi lors de l'étude rétrospective.

II. Généralités

1. Définition de la déficience intellectuelle

Lors de la conférence de l'Organisation Mondiale de la Santé (OMS), en Roumanie en novembre 2010, portant sur les enfants et les jeunes avec DI, deux points importants ont été mis en avant :

- la nécessité de distinguer la DI légère (DIL ; QI=50-69) et la DI sévère (DIS ; QI<50) dans l'estimation de la fréquence de la DI, la DIS étant 2 à 6 fois moins fréquente que la DIL ;
- le constat d'un défaut de connaissances sur le nombre de personnes atteintes, ce qui rend d'autant plus difficile l'appréciation des besoins et de la qualité des soins pour les personnes avec DI (OMS, rapport de 2010).

Les batteries les plus utilisées en France évaluant le QI sont les échelles de Wechsler adaptées selon l'âge (WAIS–IV, WISC–IV, WPPSI–IV). Elles sont destinées à évaluer l'intelligence des enfants principalement d'âge scolaire. Avant l'âge de cinq ans, on qualifie plutôt les troubles du développement présentés comme un retard de développement, un retard psychomoteur ou un retard des acquisitions (RA).

2. Prévalence de la déficience intellectuelle, en France et dans le monde

On considère, lors de la mesure de la prévalence de la DI, uniquement les personnes avec DI « fixée », c'est-à-dire stable dans le temps et ayant débuté avant l'âge de 18 ans, même si le fonctionnement général de ces personnes peut être par la suite influencé par l'environnement dans lequel elles vivent.

On exclut donc les troubles intellectuels dont l'étiologie se réfère à des maladies survenues après l'âge de 18 ans (exemple : neurodégénératives comme Alzheimer), ou à des accidents (traumatisme crânien suite à un accident de la voie publique). Il faut aussi souligner que, selon l'étiologie de la DI et sa prise en charge, elle peut s'aggraver au fil du temps.

A partir de données de la population générale, notamment celles des registres de handicap de l'enfant, (RHEOP : Registre des Handicaps de l'Enfant et Observatoire Périnatal de l'Isère et des deux Savoie) la prévalence de la DIS en France en 2010 est estimée autour de 3 pour 1000 enfants résidents, à l'âge de 7 ans, dans les départements couverts par ces registres.

Pour la prévalence de la DIL en France, il faut regarder les données d'enquêtes, effectuées soit en population générale, soit sur des groupes à risque. En population générale, la prévalence était estimée à 18 pour 1000, en incluant les DIL dissociées et les DIL « limites », soit un QI compris entre 70 et 74 (David *et al*., 2014).

Dans la littérature internationale, beaucoup d'études ne distinguent pas les DIL et les DIS. Une revue ancienne de la littérature, mais importante (Roeleveld and Zielhuis, 1997) montrait que la prévalence totale de la DI, en population générale mondiale, se décomposait en 3.8 pour 1000 pour la DIS et jusqu'à 30 pour 1000 pour la DIL. Une méta-analyse portant sur 52 études menées entre 1980 et 2009 conforte cette estimation, avec un taux de prévalence de la DI (DIL et DIS) calculé à 10,4 pour 1000 [9,6-11,2] (Maulik *et al*., 2011)

En résumé, pour les DIL, on peut retenir un taux de prévalence actuel variant entre 10 et 20 pour 1000 en France comme dans les autres pays développés. Pour les DIS, on retiendra un taux de prévalence de 3 à 4 pour 1000, qui est un taux stable dans le temps en France comme à l'étranger.

3. Causes de la déficience intellectuelle

L'identification de l'étiologie de la DI est primordiale afin de répondre aux questions des familles et d'optimiser la prise en charge des patients. Elle permet notamment d'éviter de passer à côté d'une cause curable (exemple de la phénylcétonurie, pathologie d'origine génétique mais pouvant être traitée efficacement) et d'évaluer le risque de transmission aux apparentés.

Devant l'extrême hétérogénéité clinique et génétique des DI, cette recherche de l'étiologie causale reste difficile et souvent sans résultat concret. S'il est admis que la DI est la conséquence d'un évènement qui perturbe le développement cérébral en prénatal, en périnatal ou en postnatal, les causes de la déficience intellectuelle sont multiples et les facteurs souvent intriqués. Il peut d'agir de causes d'origines environnementales ou génétiques, héréditaires ou acquises.

3.1 Facteurs psychosociaux et économiques

Le niveau socio-économique (incluant le contexte économique et le niveau d'éducation des parents) joue un rôle certain sur la prévalence de la DIL. Plusieurs études ont démontré un lien inverse entre ces 2 facteurs : la prévalence de la DIL est plus basse lorsque le niveau socio-économique est plus élevé (David *et al*., 2014; Leonard and Wen, 2002). Il a été établi une corrélation entre un niveau bas d'éducation maternelle et le risque de déficience intellectuelle chez l'enfant, pour la DIL comme pour la DIS (Croen *et al*., 2001).

Plus récemment, ce lien entre contexte socio-économique défavorable (lieu d'habitation en zone défavorisée, revenu faible) et prévalence de la déficience intellectuelle a été de nouveau mis en évidence (Emerson, 2012) avec un risque plus grand pour les enfants présentant une DIL d'être exposés à des conditions sociales défavorables dans le futur. En revanche, la prévalence de la DIS varie peu selon le milieu socio-économique (Leonard and Wen, 2002).

Pour résumer, plusieurs études s'accordent à suggérer un impact des facteurs psychosociaux et économiques (stress maternel, statut socio-économique de la famille, maltraitance) sur la survenue d'un déficit intellectuel.

3.2 Facteurs environnementaux, quelques exemples

3.2.1 Exemple d'exposition à un risque : Troubles liés à l'alcoolisation fœtale

Parmi les facteurs environnementaux, on considère les troubles causés par l'alcoolisation fœtale regroupant les manifestations qui peuvent survenir chez un individu dont la mère a consommé de l'alcool durant la grossesse. L'atteinte cérébrale fœtale fait toute la gravité de ces symptômes.

L'exposition maternelle à l'alcool pendant la grossesse engendre un grand nombre de pathologies cliniques allant de la forme la plus caractéristique et la plus sévère, le syndrome d'alcoolisation fœtale (SAF), à des formes incomplètes se traduisant par un retard développemental et/ou un trouble d'adaptation sociale, réunies sous le terme de *Fetal Alcohol Spectrum Disorder* (FASD).

Selon les critères diagnostiques établis par les sociétés pédiatriques (Hoyme *et al.*, 2005) le SAF comporte :

• une dysmorphie faciale parfois difficile à mettre en évidence (comprenant des fentes palpébrales raccourcies, un philtrum lisse et une lèvre supérieure mince) ;

• un retard de croissance (taille et/ou poids) prénatal et/ou postnatal;

• des anomalies des structures cérébrales (témoignant de troubles survenus durant la morphogénèse) et/ou une microcéphalie

Les anomalies du système nerveux sont directement liées aux effets de l'alcool et leurs conséquences s'expriment de manière variable avec l'âge. La forme clinique la plus fréquente est une des formes partielles de SAF, qui est responsable de troubles neuro-développementaux, d'échecs scolaires et de troubles du comportement, notamment à l'adolescence.

L'incidence du SAF en France est estimée de 0.17 à 0,5 pour 1000 naissances (Bloch *et al*., 2008). Celle de l'ensemble des troubles liés à l'alcoolisation fœtale seraient de 2,3 à 6,3 pour 100 naissances en incluant tous les pays de l'Europe de l'Ouest et les Etats-Unis (May *et al*., 2011).

3.2.2 Exemple d'évènement inattendu : Prématurité

La DI est plus fréquemment observée parmi les enfants nés prématurés. Dans la population ciblée des grands prématurés, à moins de 33 SA (semaines d'aménorrhée), l'étude Epipage (Etude épidémiologique sur les petits âges gestationnels) a mis en évidence une prévalence de la DI (DIS et DIL) estimée à 12%, soit 4 fois plus élevée que les 3% [1,7-6,1] observés dans le groupe contrôle des enfants nés à terme (entre 39 et 40 SA) (Larroque *et al.*, 2008).

Il est intéressant de noter aussi que les écarts sont à peine diminués après ajustement sur le statut socio-économique. La diminution de l'âge gestationnel augmente donc le risque d'une altération de l'efficience cognitive globale.

3.3 Origine génétique

Pour ce qui est des causes génétiques de la DI, elles représenteraient 15 à 50% de toutes les étiologies identifiées de DI (Srour and Shevell, 2014) ; tous les modes de transmission mendéliens (autosomique ou lié au chromosome X, dominant ou récessif) et non mendéliens (c'est-à-dire sans altération de la séquence du génome nucléaire) sont décrits.

3.3.1 Exemple d'aneuploïdie : La trisomie 21

La trisomie 21 (ou syndrome de Down) est la plus connue et la plus fréquente des aneuploïdies autosomiques. Elle se définit par la présence surnuméraire, en partie ou en totalité, d'un troisième exemplaire du chromosome 21, qui s'explique le plus souvent par une non-disjonction chromosomique lors de la méiose dans les gamètes parentaux.

Le risque d'anomalies dans la méiose augmente avec l'âge maternel, ce qui en fait le principal facteur de risque de survenue de la trisomie 21. De plus, si l'un des deux parents est porteur d'un remaniement de la structure chromosomique, telle une translocation robertsonienne entre les chromosomes 14 et 21, cela augmente la fréquence des anomalies méiotiques, et donc induit un risque accru pour les futurs fœtus de porter cette trisomie. Une grossesse antérieure avec un fœtus atteint de trisomie 21 représente aussi un facteur de risque pour le fœtus à venir.

Dans ce contexte, on observe en moyenne 27 grossesses sur 10 000 porteuses de cette anomalie génétique (données de la Haute Autorité de Santé (HAS) de 2010, hors pertes fœtales spontanées) et cette fréquence augmente avec l'âge maternel. En France, en 2005, la prévalence à la naissance était estimée à 5,1 pour 10 000 naissances et le taux d'interruption médicale de grossesse après diagnostic anténatal était de 78% (Rousseau *et al.*, 2010).

Ces enfants atteints présentent des particularités morphologiques caractéristiques, tels un visage rond, des fentes palpébrales orientées en haut et en dehors, un épicanthus, une nuque plate, ou encore un pli palmaire unique bilatéral. Ils présentent aussi des malformations, notamment cardiaques (canal atrio-ventriculaire), digestives (atrésie duodénale), neurologiques (épilepsie, apnées du sommeil), etc. qui peuvent être à l'origine de complications sérieuses et nécessitent un suivi médical adapté.

La déficience intellectuelle est variable, souvent classée parmi les DIL, et s'accompagne d'une hypotonie musculaire et d'une laxité articulaire quasi constantes. Un projet coordonné, éducatif, rééducatif et social doit être mis en place pour faciliter l'intégration de la personne trisomique (en milieu ordinaire ou adapté) et son épanouissement dans la société.

La prévalence de la trisomie 21 a diminué significativement dans plusieurs pays après la mise en place du dépistage prénatal. L'objectif de ce dépistage est de donner à la femme enceinte ou aux couples le souhaitant, une information éclairée sur le risque de trisomie 21 de leur fœtus, afin de leur permettre de décider librement de la poursuite ou non de la grossesse si la trisomie est diagnostiquée. Il n'est pas obligatoire mais doit être systématiquement proposé lors d'une nouvelle grossesse. Celui-ci se divise en plusieurs étapes en fonction du calendrier relié à la grossesse.

En première intention, les recommandations de l'HAS parues en 2009 préconisent un dépistage combiné du 1er trimestre, à réaliser entre 11 SA et 13 SA + 6 jours. Il associe la mesure échographique de la clarté nucale et le dosage des marqueurs sériques du 1er trimestre (dosage de βHCG et PAPP-A dans le sang maternel). Pour les femmes n'ayant pu bénéficier du dépistage combiné du 1er trimestre pour des raisons de délais notamment, il peut être proposé un dépistage par les marqueurs du 2^{e} trimestre (dosage de hCG totale et de l'alphafœtoprotéine). D'après ces informations, on établit un résultat sous forme de risque de trisomie 21.

Si le risque de trisomie 21 est estimé égal ou supérieur à 1/250 après dosage des marqueurs, une confirmation diagnostique par caryotype fœtal doit être proposée. Selon l'avancée de la grossesse, l'examen invasif réalisé est une choriocentèse (à partir de 11 SA) ou une amniocentèse (à partir de 15 SA) dans le but d'extraire des villosités choriales ou du liquide amniotique de l'ADN fœtal pour effectuer un caryotype. De plus, si la clarté nucale est mesurée égale ou supérieure à 3,5 mm, qu'il

existe d'autres signes d'appels échographiques, ou si le risque est supérieur ou égal à 1/50 d'emblée, un caryotype fœtal est également proposé.

Récemment, des tests ADN libre circulant de la trisomie 21 (ADNlcT21) ont été introduits dans le dépistage anténatal de la trisomie 21. L'objectif de ces tests est d'évaluer, à partir de l'ADN fœtal que l'on sait présent dans le sang maternel pendant la grossesse, la proportion relative de chacun des chromosomes 13, 18 et 21. Ceci permet de mettre en évidence un « surplus » de matériel chromosomique, qui est observé lorsque le fœtus est porteur d'une trisomie 13, 18 ou 21.

Il s'agit donc, pour les patientes qui souhaitent bénéficier de ce test, de réaliser une simple prise de sang (sur tube « Streck », contenant un conservateur particulier), dans le cadre d'un Dépistage Prénatal Non Invasif (DPNI). Depuis 2016, ce test s'inscrit dans les recommandations de pratiques professionnelles de l'HAS.

Il peut être réalisé à partir de 10 SA et dispose d'une très bonne sensibilité. Il est proposé à toutes les femmes enceintes dont le niveau de risque de trisomie 21 fœtale est compris entre 1/1000 et 1/51 à l'issue du dépistage combiné du 1er trimestre. D'autres indications peuvent être retenues pour la réalisation d'un DPNI : l'âge maternel égal ou supérieur à 38 ans pour les patientes n'ayant pas pu bénéficier du dépistage par des marqueurs sériques maternels, l'un des membres du couple est porteur d'une translocation robertsonienne impliquant le chromosome 21, ou encore un antécédent de grossesse avec aneuploïdie.

Il est aussi systématiquement proposé pour les grossesses gémellaires, car l'estimation du risque de trisomie 21 par le dosage des marqueurs sériques n'est pas fiable. Il n'est pas recommandé actuellement pour le dépistage d'autres anomalies chromosomiques, comme les anomalies des chromosomes sexuels ou les syndromes microdélétionnels, du fait de difficultés d'interprétation des résultats du test.

Malgré leur technologie innovante indéniable, ces tests ne se substituent pas à l'ensemble des tests proposés dans le cadre du dépistage de la trisomie fœtale. Il est impératif de réaliser un suivi échographique et d'effectuer une confirmation diagnostique. En effet, en cas de résultat positif d'un test ADNlcT21, le diagnostic doit être confirmé par la réalisation d'un caryotype fœtal.

A terme, ces tests pourraient contribuer à l'amélioration des performances du dépistage anténatal de la trisomie 21, à diminuer le taux de faux-positifs, et donc à réduire le nombre d'indications pour un examen invasif à visée diagnostique. De plus, avec un diagnostic obtenu plus précocement, ils pourraient permettre de diminuer le nombre d'interruptions médicales de grossesses tardives, dont les conséquences psychologiques peuvent être lourdes.

3.3.2 Exemple d'une maladie liée à une anomalie de répétition de triplets nucléotidiques : le syndrome de l'X fragile

Le syndrome de l'X fragile est une maladie génétique rare, de transmission dominante liée à l'X, à pénétrance incomplète chez les filles. On estime sa prévalence à 1 sur 2500 (pour la mutation complète) à 1 sur 4000 (cas symptomatiques) pour les deux sexes. Cette pathologie est due à une expansion anormale d'un triplet nucléotidique CGG (cytidine, guanine, guanine) dans le locus Xq27.3, aux abords du gène *FMR1,* dans la région 5' non traduite. Cette anomalie induit une hyperméthylation de la région du promoteur de ce gène, ce qui le rend « silencieux », c'est-à-dire que ce phénomène induit une réduction de la production de la protéine FMRP (*Fragile X mental retardation protein*). Celle-ci est impliquée, entre autres, dans la plasticité synaptique et la signalisation dendritique.

On considère une mutation complète du gène *FRM1* à partir de plus 200 répétitions de triplets CGG. Les personnes atteintes présentent un tableau clinique variable. La plupart des enfants présentent un retard de développement, notamment au niveau du langage et de la marche, ainsi qu'un déficit intellectuel identifiable dans l'enfance, variable également. On peut observer aussi des troubles du comportement, avec une hyperkinésie, une humeur instable, une anxiété, voire des troubles plus sévères, de type autistique. Ces caractéristiques sont peu spécifiques de ce syndrome et sont plus marquées chez les garçons que chez les filles. Les signes physiques restent discrets.

Ces allèles mutés proviennent de prémutations, c'est-à-dire d'allèles instables, prémutés (55 à 200 répétitions de CGG), transmis à la génération suivant avec le risque d'une amplification de l'expansion des triplets. Les personnes possédant un ou deux allèles prémutés sont dits « porteuses » de la prémutation. Les femmes porteuses ont donc un risque accru d'avoir des enfants porteurs de la mutation complète et atteints du syndrome de l'X fragile. Les hommes, en revanche, ne transmettent pas d'allèles prémutés ou mutés à leurs fils (puisqu'ils transmettent le chromosome Y) mais à toutes leurs filles.

Ces individus porteurs de la prémutation sont, pour la majorité, asymptomatiques. Mais avec l'âge, ils peuvent développer une maladie dégénérative rare, qui affecte principalement les hommes, avec un risque cumulé dans la population générale de 1 sur 8000. C'est le syndrome tremblement-ataxie, lié à une prémutation de l'X-fragile. Il se caractérise par la survenue à l'âge adulte (le plus souvent après 50 ans) d'un tremblement intentionnel et d'une ataxie cérébelleuse. Les personnes atteintes font des chutes fréquentes et peuvent souffrir de neuropathies périphériques et d'une dysfonction du système autonome (incluant incontinence, hypotension orthostatique…) (Berry-Kravis *et al.*, 2007).

3.3.3 Exemple de transmission non mendélienne : le phénomène « d'empreinte » parentale

La découverte d'un phénomène de marquage des génomes parentaux, appelé « empreinte », a permis de mieux caractériser des syndromes connus mais dont le mécanisme génétique était mal compris jusque-là (Chamberlain and Lalande, 2010). L'identification de gènes spécifiques soumis à cette empreinte parentale a mis en évidence l'impact d'une expression monoallélique (maternelle ou paternelle) de gènes spécifiques dans ces pathologies, alors qu'en situation normale, les allèles des deux parents sont actifs. Il s'agit principalement de méthylations de l'ADN, d'acétylations et de méthylations des histones qui sont à l'origine de ce phénomène de marquage épigénétique.

Les syndromes de Prader-Willi et d'Angelman sont deux exemples probants de DI résultant d'anomalies liées à l'empreinte génomique, et qui résultent d'un mode de transmission non mendélien (Gurrieri and Accadia, 2009). Le phénotype attaché au syndrome de Prader-Willi comporte une hypotonie néonatale majeure, une hyperphagie pouvant engendrer une obésité morbide, des traits dysmorphiques particuliers ainsi qu'une déficience intellectuelle globalement légère voire limite et des troubles du comportement. Les patients présentant un syndrome d'Angelman souffrent de troubles neurologiques, tels une ataxie et des crises d'épilepsie, avec une déficience intellectuelle sévère, une absence de langage et une jovialité immotivée.

Ces deux syndromes sont causés par des anomalies géniques dans la région 15q11 – 15q13, qui contient plusieurs gènes soumis à empreinte parentale. Ces anomalies peuvent être multiples : délétions chromosomiques de taille importante (70 à 75% des cas), disomies uniparentales (25% de disomies maternelles sont à l'origine du syndrome de Prader-Willi), mutations diverses dans la région concernée, etc. Le syndrome de Prader-Willi est causé par la perte de l'expression de plusieurs gènes soumis à empreinte, et qui se situent normalement sur le chromosome 15 d'origine paternelle, tandis que le syndrome d'Angelman est lui causé par la suppression de l'expression du gène *UBE3A*, normalement actif sur le chromosome 15, cette fois d'origine maternelle.

4. Explorations des causes génétiques de la DI

Depuis l'établissement de premier caryotype en 1959, les techniques de génétique n'ont cessé d'évoluer. Elles facilitent la découverte de nouveaux gènes impliqués dans des maladies génétiques, permettent l'identification de certaines mutations autrefois non détectées et offrent ainsi de nouvelles possibilités en matière de diagnostic étiologique et de recherche clinique.

4.1 L'ACPA (Analyse Chromosomique sur Puce à ADN)

Depuis le milieu des années 2000, le caryotype « standard » n'est plus l'examen de première intention pour l'exploration d'anomalies chromosomiques associées à une déficience intellectuelle. Il est remplacé par des techniques moléculaires, pangénomiques, qui s'appuient sur le principe de l'hybridation génomique comparative (CGH : *Comparative Genomic Hybridization*), en utilisant des puces : c'est l'Analyse Chromosomique sur Puce à ADN, également appelée caryotype moléculaire.

Ces puces permettent d'obtenir une résolution de quelques centaines de paires de bases, soit 10 à 500 fois supérieure à celle du caryotype morphologique (3 à 5 Mb) et ainsi de détecter des déséquilibres chromosomiques non visibles sur le caryotype standard. La technique de CGH se base sur la comparaison de deux ADN génomiques, en quantité équimolaire et marqués par un fluorochrome différent, hybridés sur la même puce.

La fluorescence de ces marqueurs est ensuite mesurée dans un scanner de puces, qui numérise les données et calcule un rapport d'intensité de fluorescence entre les 2 ADN. Le logiciel qui s'occupe de l'extraction de ces données génère un résultat sous forme de représentation graphique de ces rapports de fluorescence chromosome par chromosome, rapports qui sont déséquilibrés dans le cadre de la découverte d'un CNV (*Copy Number Variation*). Par définition, un CNV est une variation quantitative du génome d'une personne dans une région chromosomique donnée, qu'il s'agisse d'un gain ou d'une perte de matériel, par rapport à un génome de référence.

Cependant, en dépit des connaissances acquises sur le génome à l'heure actuelle et de la forte résolution de cette analyse, il n'est pas toujours possible de conclure sur la pathogénicité d'un CNV. Il devient alors nécessaire d'étudier plus finement le génome au niveau moléculaire pour tenter d'élucider l'étiologie de la DI présentée par le patient. Ceci justifie l'intérêt de développer de nouvelles techniques plus performantes permettant d'identifier des variations géniques à l'échelle nucléotidique.

4.2 Le séquençage haut débit

Pour décrypter le code génétique que représente l'ADN, le biochimiste Frederick Sanger a mis au point en 1977 un procédé chimique permettant d'établir l'ordre linéaire des différentes bases nucléotidiques, dans une structure macromoléculaire et de reconstituer leurs enchaînements (Sanger *et al.*, 1977). Ce procédé, appelé séquençage Sanger, se base sur la reconstitution d'un brin d'ADN complémentaire à celui dont on veut déterminer la séquence, grâce à une réserve de nucléotides libres et à une ADN polymérase, présentes dans le milieu réactionnel.

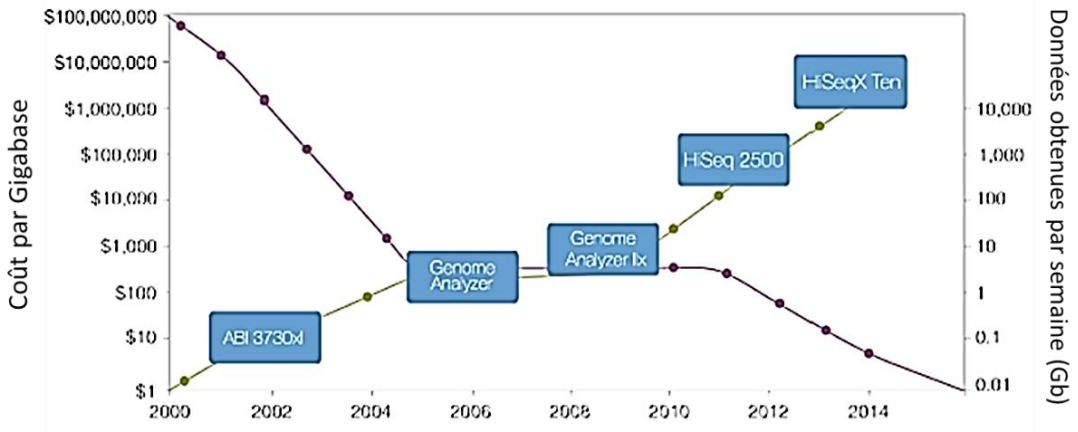


Figure 1. Evolution du coût du séquençage du génome et de la quantité de données recueillies depuis 2000

Initialement, le séquençage d'échantillons d'ADN s'effectuait de façon manuelle, puis devint automatisé, avec notamment l'utilisation de séquenceurs automatiques capillaires dans le milieu des années 90. Depuis 2006, grâce aux innovations technologiques et informatiques, des séquenceurs à très haut débit ou séquenceurs de nouvelle génération (Bai *et al.*, 2005), peuvent séquencer simultanément et de manière indépendante jusqu'à 100 Gb d'ADN. Ces nouvelles technologies de séquençage massif sont regroupées sous l'acronyme NGS pour *Next Generation Sequencing*.

Il est intéressant de constater que, par rapport au tout premier séquençage du génome humain qui fut obtenu en 15 ans (International Human Genome Sequencing Consortium *et al*., 2001) et qui coûta près de 3 milliards de dollars, des séquenceurs haut débit comme le HiSeq X® Ten mis sur le marché en 2014, peuvent aujourd'hui séquencer 45 génomes en un seul jour, pour environ 1000 dollars chacun.

Grâce à l'essor des techniques moléculaires, on estime actuellement qu'il est possible d'identifier le gène responsable de la DI dans 55 à 70 % des cas pour des patients présentant une DI modérée ou sévère (Vissers *et al.*, 2016).

L'outil que constitue l'analyse de l'exome aujourd'hui est une innovation majeure dans la recherche d'anomalies ponctuelles à l'échelle nucléotidique. Le séquençage de l'exome concerne les exons ainsi que les régions introniques flanquantes, soit moins de 2% du génome, ce qui représente environ 62 Mb lues en une seule fois. Jusqu'à présent, plus de 1200 gènes ont été mis en évidence dans la DI et c'est aujourd'hui vers l'analyse de l'exome que se porte la recherche pour découvrir de nouveaux gènes.

A ce jour, seul le séquençage de l'exome est accessible à l'échelle des laboratoires Hospitaliers et Universitaires, car celui du génome entier demeure encore trop coûteux en termes d'équipements et de matériel informatique, pour l'analyse et le stockage des données.

III. Matériels et Méthodes

1. Objectif principal et plan des deux études

Le séquençage de l'exome a pour but princeps d'identifier la cause de la déficience intellectuelle du patient, par recherche de mutations nucléotidiques et recherche de CNV de très petites tailles, non décelés par le caryotype moléculaire.

Pour ce projet, nous effectuerons deux types d'études en parallèle. De manière rétrospective, nous analyserons des données des exomes obtenus lors de différentes analyses, dont celles obtenues dans le cadre du projet HUGODIMS, formant une cohorte de patients présentant tous une DI, isolée ou syndromique, pour définir la meilleure stratégie d'analyse informatique pour identifier de nouveaux variants géniques, impliqués dans la DI.

En effet, actuellement, il n'existe pas de méthode standardisée pour l'analyse informatique des données obtenues avec la technique de l'exome. Nous allons donc utiliser des données issues du séquençage de ces exomes rétrospectifs afin d'établir une stratégie performante de tri des variants d'intérêt. Nous essaierons en particulier de définir un score qui prendra en compte de nombreux paramètres (notamment les prédictions informatiques sur les mécanismes d'épissage et sur l'aspect structure/fonction des protéines, les données phénotypiques, la fréquence allélique, la présence dans les différentes bases de données publiques).

La puissance du score sera évaluée sur sa capacité à déceler rapidement le ou les variants pathogènes impliqués dans le phénotype du patient. Le laboratoire dispose déjà de 21 données d'exomes qui serviront de contrôles positifs avec une mutation causale déjà identifiée. Le score de tri de variants que nous retiendrons devra, sur ces données d'exomes contrôles, permettre l'identification du variant pathogène dans 100% des cas.

Par ailleurs, parmi l'ensemble des cas rétrospectifs que nous analyserons, nous nous attendons à trouver un variant probablement pathogène sur un gène répertorié dans la base de données OMIM (*Online Mendelian Inheritance in Man*) dans 20 à 30% des cas (Lee *et al*., 2014). Cette base de données correspond à un catalogue de toutes les pathologies connues avec composante génétique, reliées pour la plupart au gène impliqué.

La seconde partie de ce projet consiste en une étude prospective de patients issus des consultations de génétique du CHU de Poitiers par analyse de l'exome. En pratique, les expérimentations se diviseront en 4 réactions de séquençage, comprenant 12 échantillons de patients chacun, soit un total de 48 exomes à analyser. De plus, nous réaliserons aussi l'analyse de 30 exomes de patients issus d'expérimentations antérieures et dont l'interprétation biologique est en attente.

Pour cette analyse, nous utiliserons le protocole informatique optimisé dans la première partie du projet pour l'analyse de l'exome de cette cohorte prospective, dont la technique de séquençage aura été effectuée au Laboratoire de Génétique Biologique du CHU de Poitiers.

A l'issue de cette étude globale, nous aurons deux types de situations concernant nos patients :

(1) un variant considéré pathogène situé sur un gène OMIM ; dans ce cas, le résultat sera rendu au clinicien après une réunion de concertation clinico-biologique et pourra être utilisé pour le conseil génétique.

(2) un ou plusieurs variants probablement pathogènes dans des gènes non répertoriés dans OMIM. Ces variants classés « Recherche » feront l'objet d'une discussion détaillée dans ce rapport et seront déposés dans la base de données GeneMatcher (https://www.genematcher.org/), qui vise à établir des cohortes génotype/phénotype internationales de patients ayant bénéficié d'analyses pangénomiques.

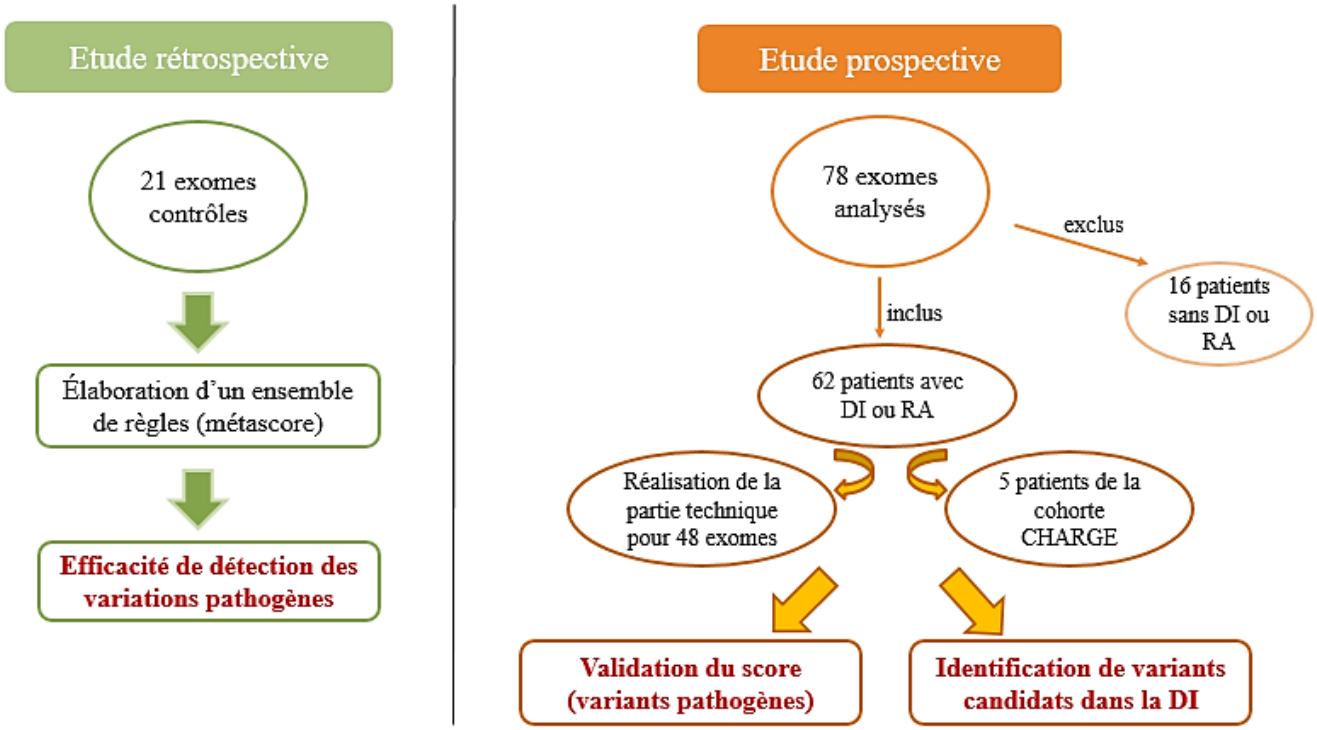


Figure 2. Plan des deux études

2. Patients analysés dans l'étude prospective

Pour l'étude prospective, ces patients doivent présenter une déficience intellectuelle et/ou une anomalie du développement dont l'étiologie demeure inconnue. Ils auront eu au préalable un bilan d'analyses classiques de génétique, fait au Laboratoire de Génétique Biologique du CHU de Poitiers : analyse de l'X-fragile, caryotype moléculaire et/ou analyse d'un panel de 275 gènes les plus fréquemment impliqués dans la déficience intellectuelle, dérivé de celui de Piton et collaborateurs (Redin *et al.*, 2014). La prescription d'une analyse de l'exome est faite par un médecin généticien lors d'une consultation spécialisée et s'accompagne obligatoirement de la signature d'un consentement éclairé par le patient.

Parmi ces échantillons, nous incluons aussi des patients présentant un phénotype clinique très évocateur d'un syndrome CHARGE (acronyme anglais de ***C****oloboma,* ***H****eart defect,* ***A****tresia of the choanae,* ***R****etarded growth and development,* ***G****enital hypoplasia,* ***E****ar anomalies/deafness* soit en français : colobome, malformations cardiaques, atrésie choanale, retard de croissance et/ou retard mental, hypoplasie génitale, anomalies des oreilles et/ou surdité) mais pour lesquels l'analyse NGS sur les gènes *CHD7, EFTUD2* et *HOXA1*, principaux gènes impliqués préalablement dans ce syndrome, ne retrouve pas de variation pathogène.

L'analyse de l'exome peut s'effectuer selon deux modes de séquençage : en simplex (patient seul) ou en trio (analyses concomitantes de l'exome du patient et de ses parents). L'analyse en trio nécessite le séquençage de 3 individus et permet d'identifier des variants présents chez le patient, non retrouvés chez les parents, dits « *de novo* ».

Dans les analyses antérieures, il est apparu que dans la plupart des cas de DI, les mutations impliquées sont celles rapportées *de novo* (Lupski, 2010), mais cette stratégie, malgré son intérêt certain dans le tri des variants, présente un coût important pour le laboratoire. Ainsi, pour des raisons économiques, nous avons opté dans un premier temps pour un séquençage des patients en simplex.

3. Technique de l'exome

3.1 Extraction de l'ADN

Tout d'abord, nous extrayons l'ADN des patients à partir de sang total, obtenu par prélèvement sanguin sur tube EDTA, grâce au kit MagPurix Blood DNA Extraction®, de Zinexts. Il s'agit d'une extraction automatique avec l'automate MagPurix® (Zinexts Life Science Corporation) à partir d'un

échantillon de sang total. Cet automate utilise une technologie de traitement fluidique basée sur la capture des molécules d'ADN par des nanoparticules magnétiques.

3.2 Contrôle de la concentration et de la pureté

La quantité d'ADN est cruciale pour l'expérimentation. Post-extraction, nous dosons la concentration d'ADN avec le NanoDrop®, un spectrophotomètre à micro-volume, ainsi qu'avec le Qubit™, un fluorimètre associé avec des réactifs spécifiques pour la quantification spécifique d'ADN ou d'ARN. L'objectif est de valider la qualité et la quantité de ces ADN après l'extraction pour la construction de librairies NGS.

Contrairement au Nanodrop®, le Qubit™ permet de mesurer de très faibles concentrations d'ADN. Un autre avantage du Qubit™ est sa spécificité : les réactifs utilisés se fixent spécifiquement à l'ADN d'intérêt, les concentrations obtenues ne sont donc pas affectées par la présence de contaminants. Le Qubit™ ne permet donc pas la détection de contaminants présents dans les ADN extraits.

Cependant, avant d'effectuer une technique de séquençage, il est important d'évaluer la contamination par des protéines dans une solution d'acides nucléiques, c'est pourquoi nous utilisons le ratio A260 nm/A280 nm obtenu par le NanoDrop®. De plus, il n'y a pas besoin, contrairement au Qubit™, d'une préparation préalable des échantillons avant la mesure des concentrations. La pureté d'une solution d'acides nucléiques est considérée comme acceptable lorsque le ratio A260/A280 est compris entre 1,8 – 2,0 pour l'ADN.

Pour la technique, nous avons besoin d'une concentration d'ADN de 25 ng/µL. Pour atteindre cette concentration de manière précise, nous effectuons des dilutions successives à 100 ng/µL puis à 25 ng/µL avec de l'eau sans nucléase, dont nous vérifions à nouveau la concentration d'ADN avec le Qubit™.

Pour les étapes suivantes, nous utilisons les différents réactifs et les tampons fournis dans les kits de réactifs SureSelect^{QXT} d'Agilent Technologies® adaptés à la plateforme NextSeq d'Illumina dont nous disposons.

3.3 Préparation des librairies d'ADN à séquencer

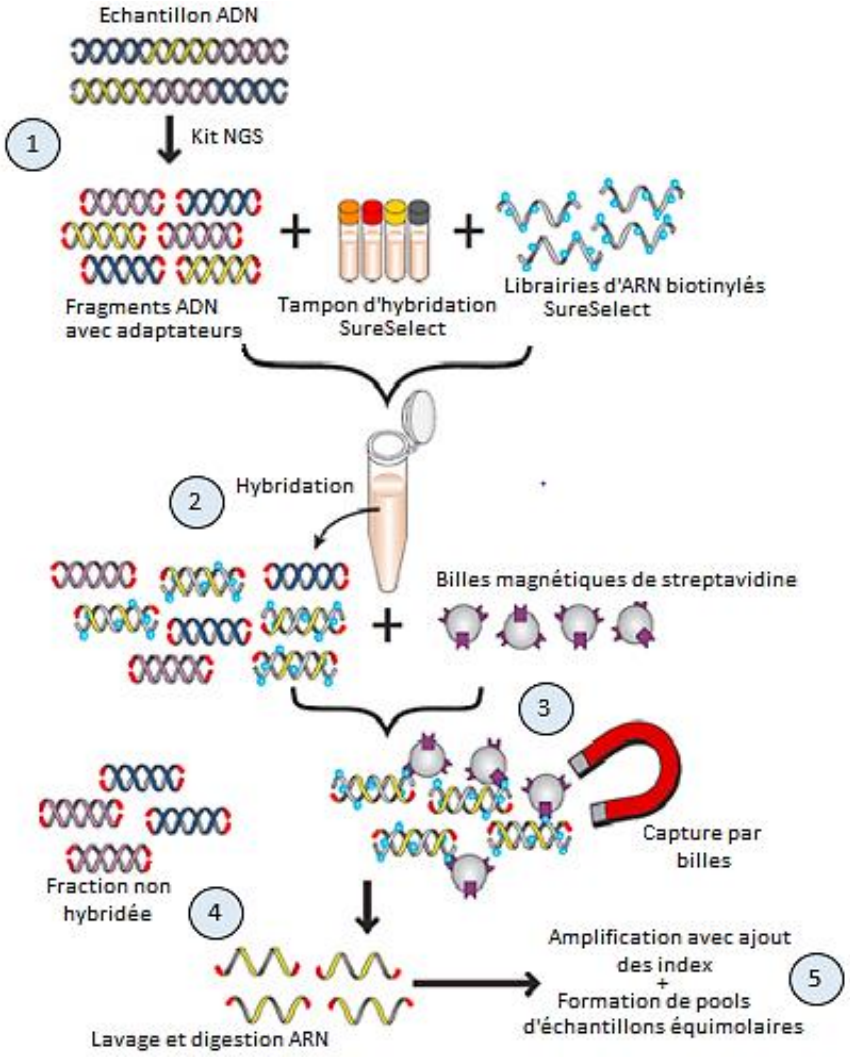


Figure 3. Etape 1 de l'analyse du l'exome : Préparation des librairies (*d'après des images d'Agilent Technologies*®)

1. Fragmentation des échantillons d'ADN et fixation d'adaptateurs aux extrémités (purification et amplification, non incluses sur le schéma)
2. Hybridation des fragments d'ADN avec les librairies fournies dans le kit Agilent SureSelect, composées d'oligonucléotides biotinylés
3. Capture des hybrides d'ADN à l'aide des billes magnétiques de streptavidine qui se fixent uniquement aux hybrides biotinylés
4. Elimination des fragments non hybridés et lavage des fragments capturés
5. Amplification des librairies capturées, avec ajout de marqueurs index (purification non incluse sur le schéma) et préparation du pool d'échantillons pour séquençage multiplexe

3.3.1 Addition des fragments et des adaptateurs

La préparation des librairies est effectuée en utilisant les technologies de fragmentation (via des transposases) de type SureSelect d'Agilent Technologies® (étape n°1 sur la Figure 3). Pour des

raisons techniques, la préparation des librairies s'effectue par tranches de 8 échantillons. Après la fragmentation de l'ADN, l'ajout d'adaptateurs est réalisé aux extrémités générées.

Pour ce faire, nous effectuons un mélange dans des tubes en barrette composé de 17 µL de SureSelect Buffer (préalablement décongelé), de 2 µL de SureSelect QXT Enzyme Mix, et de 2 µL d'ADN dilué (un tube par échantillon ADN). Après avoir homogénéisé la solution, les tubes sont fermés, agités rapidement avec un agitateur de type Vortex, puis centrifugés rapidement.

Nous procédons ensuite à la fragmentation de l'ADN, à l'aide d'un thermocycleur programmé pour une incubation de 10 minutes à 45°C puis 1 minute à 4°C. Nous rajoutons ensuite 32 µL de 1XSureSelect QXT Stop Solution pour arrêter la réaction. Les échantillons sont agités rapidement et sont incubés à température ambiante pendant 1 minute.

3.3.2 Purification des librairies-adaptateurs

Un lavage à l'AMPure XP (contenant des billes magnétiques) est alors effectué pour purifier les fragments obtenus, et ne garder que les fragments contenant des adaptateurs. Nous rajoutons 52 µL d'AMPure, préalablement agités avec précaution pour disperser les billes contenues dans la solution, dans les échantillons obtenus précédemment. Après une agitation rapide de cette nouvelle solution, nous effectuons une centrifugation courte.

Après 5 minutes d'attente à température ambiante, nous procédons à une séparation de 5 minutes sur plaque magnétique, puis le surnageant est éliminé. Nous utilisons 200 µL d'éthanol à 70% pour laver les échantillons, avec une attente d'une minute à température ambiante puis élimination du surnageant ; ce lavage est réitéré une deuxième fois dans les mêmes conditions. Il faut alors sécher les tubes par évaporation à 37°C (les bouchons seront préalablement enlevés), pendant environ 2-3 minutes. Nous ajoutons 13 µL d'eau sans nucléase dans chaque tube. Après une incubation de 2 minutes à température ambiante, nous effectuons une nouvelle séparation sur plaque pendant 2 minutes, puis nous transférons environ 12 µL dans de nouveaux tubes avant de les mettre dans la glace.

3.3.3 Amplification des librairies ADN (avec les adaptateurs)

Nous effectuons ensuite une amplification par PCR (PCR « pré-capture ») pour augmenter la quantité de matériel utilisable pour l'expérimentation. Pour cette étape, nous préparons un mélange réactionnel selon le Tableau 1.

Réactifs	Volume pour 1 réaction	Volume pour 9 réactions
Eau sans nucléase	25 µl	225 µl
5X Herculase II reaction Buffer	10 µl	90 µl
100mM dNTP Mix	0.5 µl	4.5 µl
DMSO	2.5 µl	22.5 µl
SureSelect QXT Primer Mix	1 µl	9 µl
Herculase II Fusion DNA Polymerase	1 µl	9 µl
Total	40 µl	360 µl

Tableau 1. Préparation du mélange pour la PCR « pré-capture »

Nous préparons de nouveaux tubes contenant 10 µL d'échantillons de librairies d'ADN purifiées, auxquels nous ajoutons 40 µL de ce mélange, pour un volume total de 50 µL. Après agitation et centrifugation rapide, nous effectuons ensuite à une amplification par PCR nommée « Pré-Capture », qui se décompose en :

- Une phase de 2 minutes à 68°C
- Une phase de 2 minutes à 98°C
- Une phase de 8 cycles de 30 secondes à 98°C, puis 30 secondes à 57°C et une minute à 72°C
- Une phase de 5 minutes à 72°C
- Une phase d'incubation finale dont la température est inférieure à 4°C.

3.3.4 Purification

Pour cette étape, nous utilisons les 50 µL d'échantillons obtenus après amplification, auxquels nous ajoutons 50 µL d'AMPure. Nous procédons ensuite à un lavage à l'éthanol, de façon identique au premier lavage effectué (voir ci-dessus paragraphe 3.3.2). Après évaporation, ajout de 13 µL d'eau sans nucléase, agitation et centrifugation rapide, on transfère 12 µL de cette solution dans de nouveaux tubes. Les tubes sont ensuite maintenus à une température de 4°C jusqu'à l'étape suivante.

3.3.5 Dosage de l'ADN des librairies

Après cette nouvelle purification à l'AMPure XP, la quantité, la qualité et la taille des fragments obtenus doivent être évaluées par un dosage avec la TapeStation® (sur un principe de séparation électrophorétique des acides nucléiques et des protéines). Dans le tube étalon, nous utilisons 3 µL de tampon D1000 et 1 µL de solution d'étalonnage. Pour les autres tubes à doser, nous utilisons la même quantité de tampon et 1 µL d'échantillon d'ADN. Puis les tubes sont agités à l'aide d'un agitateur de type MixMate® (Eppendorf) à 2000 rpm (tours par minute) pendant une minute.

Sur l'écran de l'interface informatique de l'appareil, on vérifie sur l'électrophorégramme que les pics d'ADN fragmentés sont entre 245 pb et 325 pb.

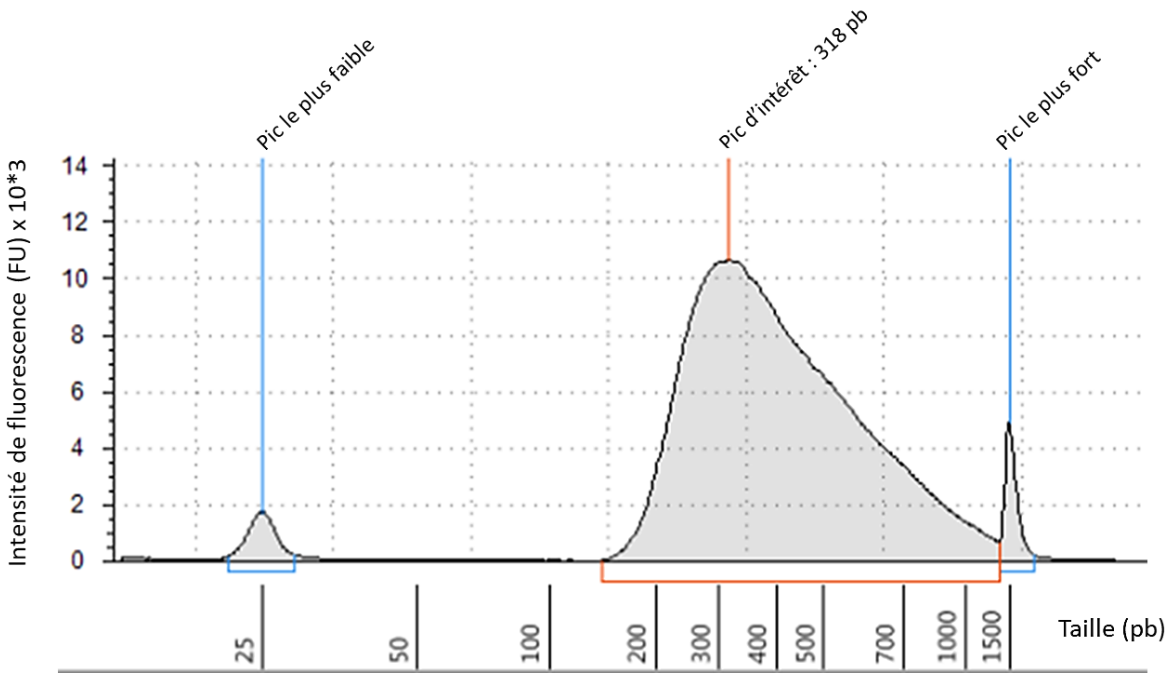


Figure 4. Exemple d'électrophorégramme obtenu avec la TapeStation® pour un échantillon d'ADN

3.4 Hybridation et capture

Une dilution des échantillons est alors nécessaire pour obtenir 125 ng/µL d'ADN (volume final de 12 µL) avant de procéder à l'hybridation.

3.4.1 Hybridation des ADN

L'étape suivante est donc l'hybridation des fragments d'ADN avec les librairies fournies dans le kit Agilent SureSelect, composées d'oligonucléotides biotinylés, qui sont les séquences d'intérêt à étudier dans le génome (ici ce sont toutes les séquences des exomes ; étape n°2 de la Figure 3).

Dans de nouveaux tubes, nous mélangeons les 12 µL d'ADN dilués avec 5 µL de mélange pré-préparé SureSelect QXT Fast Blocker Mix, puis nous incubons les tubes selon les phases suivantes:

- Une phase de 5 minutes à 95°C
- Une phase de 10 minutes à 65°C
- Une phase d'une minute à 65°C (suivie d'une pause dans le cycle, programmée sur l'automate, en attendant l'ajout du mélange décrit dans le Tableau 3)
- Une phase de 60 cycles avec 1 minute à 65°C suivie de 3 secondes à 37°C
- Une phase d'incubation finale à 65°C jusqu'à l'étape suivante

Au cours de l'étape d'hybridation, il est nécessaire de préparer la solution de RNase Block à 25% (Tableau 2) ainsi que le mélange d'hybridation-capture des librairies (Tableau 3)

Réactifs	**Volume pour 1 réaction**	**Volume pour 9 réactions**
SureSelect RNase Block	0.5 µl	4.5 µl
Eau sans nucléase	1.5 µl	13.5 µl
Total	2 µL	18 µl

Tableau 2. Préparation de la solution RNase Block à 25%

Réactifs	**Volume pour 1 réaction**	**Volume pour 9 réactions**
RNase Block solution à 25% (précédente)	2 µl	18 µl
SureSelect Capture Human All exon V7	5 µl	45 µl
SureSelect QXT Fast Hybridization Buffer	6 µl	54 µl
Total	13 µl	117 µl

Tableau 3. Préparation du mélange pour la capture des exons

Nous ajoutons 13 µL de ce mélange directement dans les tubes d'échantillons maintenus à 65°C. Après agitation et centrifugation rapide, les tubes sont à nouveau incubés pour une durée de 60 cycles de 1 minute à 65°C puis 3 secondes à 37°C.

3.4.2 Préparation des billes magnétiques de streptavidine

La capture des séquences hybridées d'exome est effectuée par des billes magnétiques de streptavidine ajoutées dans le mélange (étape n°3 de la Figure 3).

Cette étape de préparation des billes peut s'effectuer en parallèle de l'étape d'hybridation. Les billes doivent être remises en suspension vigoureusement. Pour chaque échantillon, il faut ajouter 50 µL de solution. Le lavage des billes s'effectue avec 900 µL de tampon SureSelect Binding Buffer.

Nous procédons à une séparation des billes sur une barre magnétique puis le surnageant est éliminé. Ce lavage est répété deux fois, puis nous remettons les billes en suspension dans 900 µL de tampon.

3.4.3 Capture de l'ADN hybridées avec les billes magnétiques

Dans de nouveaux tubes, nous transférons 200 µL de billes lavées et nous ajoutons 30 µL de produit d'hybridation. Ce mélange est incubé à température ambiante sous agitation pendant 30 minutes à 1800 rpm (agitateur MixMate®). Ces billes se fixent uniquement aux oligonucléotides biotinylés. Des lavages sont nécessaires pour éliminer les fragments non capturés (étape n°4 de la Figure 3).

Pour réaliser ces lavages, chaque tube contenant les billes est alors placé sur la plaque magnétique pendant environ une minute puis le surnageant du tube est jeté. Les billes sont remises en suspension par 200 µL de tampon de lavage SureSelect Wash Buffer 1. Nous réalisons à nouveau une séparation sur plaque magnétique comme précédemment avec élimination du surnageant.

Nous effectuons une nouvelle remise en suspension des billes, cette fois avec 200 µL de tampon SureSelect Wash Buffer 2 (préalablement incubés à 65°C). Les tubes contenant les billes sont alors incubés à 65°C pendant 10 minutes puis replacés sur la barre magnétique pendant une minute, le surnageant ensuite jeté.

Il faut répéter ces lavages au tampon de lavage SureSelect Wash Buffer 2 au total 3 fois, puis resuspendre in fine les billes par 23 µL d'eau sans nucléase et les conserver à 4°C jusqu'à l'étape suivante.

3.5 Indexage et traitement des échantillons pour le séquençage multiplexe

3.5.1 Amplification des librairies capturées pour l'addition des marqueurs index

Tout d'abord, nous attribuons pour chaque échantillon les index appropriés, un index étant un identifiant unique, propre à chaque séquence marquée ; pour chaque échantillon, nous disposons d'un index P7 et un index P5.

Nous effectuons alors une nouvelle amplification par PCR (PCR « post-capture ») avec le mélange préparé selon le Tableau 4.

Réactifs	Volume pour 1 réaction	Volume pour 9 réactions
Eau sans nucléase	13.5 µl	121.5 µl
5x Herculase II Reaction Buffer	10 µl	90 µl
100 mM dNTP Mix	0.5 µl	4.5 µl
Herculase II Fusion DNA Polymerase	1 µl	9 µl
Total	25 µl	225 µl

Tableau 4. Préparation du mélange pour la PCR « post-capture »

Nous ajoutons 25 µL de cette solution aux 23 µL d'échantillons d'ADN capturés, ainsi que 1 µL de Primer P7 et 1 µL de Primer P5 spécifiques à chaque échantillon (étape n°5 de la Figure 3). Le mélange réactionnel est incubé selon les phases suivantes :

- Une phase de 2 minutes à 98°C
- Une phase de 10 cycles, avec 30 secondes à 98°C, 30 secondes à 58°C et 1 minute à 98°C
- Une phase de 5 minutes à 72°C
- Une phase d'incubation finale à 4°C jusqu'à l'étape suivante.

Les billes présentes dans le milieu réactionnel sont ensuite enlevées par une séparation sur plaque magnétique (incubation de chaque tube pendant 2 minutes à température ambiante). Nous transférons 50 µL de surnageant dans un nouveau tube.

3.5.2 Purification à l'AMPure des produits de PCR « post-capture »

Pour cette étape, nous mélangeons 60 µL de solution d'AMPure XP et 50 µL d'échantillons d'ADN capturés et amplifiés obtenus précédemment. Après 5 minutes d'incubation à température ambiante, nous procédons à une séparation de 3 minutes sur plaque magnétique, puis le surnageant est éliminé.

Ensuite, nous utilisons 200 µL d'éthanol à 70% pour laver les échantillons, avec une incubation d'une minute à température ambiante puis élimination du surnageant ; ce lavage est réitéré une deuxième fois dans les mêmes conditions. L'échantillon est alors évaporé à 37°C pendant quelques minutes.

L'ADN purifié est élué par 25 µL d'eau sans nucléase. Après une incubation de 2 minutes à température ambiante, nous enlevons les billes par une séparation sur plaque magnétique (incubation pendant 2 minutes), puis nous transférons environ 24 µL d'ADN purifié dans des tubes de type LoBind (Eppendorf), qui possèdent une faible affinité pour l'ADN. Cette solution est conservée à 4°C jusqu'à l'étape suivante.

3.5.3 Dosage des librairies ADN indexés

Nous procédons alors à une vérification de la quantité des fragments d'ADN indexé par dosage avec la TapeStation®. Dans le tube étalon, nous utilisons 2 µL de tampon HS (haute sensibilité) et 2 µL de solution d'étalonnage. Pour les autres tubes à doser, nous utilisons la même quantité de tampon et 2 µL d'échantillon d'ADN. Puis nous agitons les tubes à 2000 rpm (agitateur MixMate®) pendant une minute.

Sur l'écran de l'interface numérique de l'appareil, on vérifie sur l'électrophorégramme que les pics d'ADN fragmentés sont cette fois compris entre 325 bp et 450 bp. La concentration d'ADN est contrôlée ensuite à l'aide du Qubit™.

3.6 Préparation du mélange équimolaire d'échantillons avant chargement sur le séquenceur

Pour cette étape, il est nécessaire d'avoir 12 librairies réalisées. Il faut donc effectuer deux fois toutes les étapes précédentes pour obtenir 16 librairies d'ADN. Nous mélangeons alors 12 de ces échantillons de manière équimolaire afin d'obtenir une concentration d'ADN à 14 nM dans un volume final de 30 µL. Après préparation du mélange d'échantillons, nous évaluons la concentration d'ADN total du mélange au Qubit™ puis il est incubé dans la glace. Ensuite, nous diluons cette solution de manière à obtenir une concentration finale de 2 nM avec de l'eau sans nucléase.

Nous préparons le marqueur d'ADN contrôle PhiX (PhiX library) à 4 nM, en ajoutant à 2 µL de ce marqueur, 3 µL d'une solution Tris-Chlore (pH à 8,5, avec 0,1% de Tween 20). Pour dénaturer le marqueur, nous réalisons un mélange avec 5 µL de ce PhiX et 5 µL de la solution de NaOH préparée à 0,2N. Puis nous mélangeons rapidement et nous laissons incuber à température ambiante pendant 5 minutes. Nous ajoutons ensuite 990 µL de solution d'un tampon HT1 (composant du kit « NextSeq 500/550 », préalablement décongelé).

Pour dénaturer le mélange d'échantillons, dilué à 2 nM précédemment, nous mélangeons 10 µL du mélange dilué avec 10 µL de NaOH à 0,2N. Après une centrifugation rapide, nous incubons pendant 5 minutes à température ambiante, avant d'ajouter 10 µL de Tris-HCl à 200 nM, puis 970 µL de solution préparée du tampon HT1, puis nous incubons dans la glace. Nous diluons ensuite cette solution à 1,8 pM dans du tampon HT1, pour un volume final de 1,3 mL. Une fois le marqueur contrôle PhiX et le « pool » dénaturés et dilués, nous préparons une solution commune en prenant 1,2 µL de PhiX préparé (soit 1% de la solution finale) et les 1,3 mL de la solution de mélange d'échantillons à 1,8 pM préparée précédemment.

3.7 Préparation de la cartouche de réactifs

Elle consiste en l'aspiration des solutions de 3 puits de la cartouche (voir Annexe 1), n° 20, 21 et 22, auquel on rajoute respectivement 3,9 µL d'amorces nucléotidiques de « Read 1 », 4,2 µL de « Read 2 » et 6 µL d' « Index 1 » et d' « Index 2 ». Ces solutions seront à nouveau transférées dans la cartouche au niveau des puits 7, 8 et 9 (pour respectivement les solutions des puits 20, 21 et 22).

Les « Read » et les « Index » sont des amorces nucléotidiques personnalisées fournies dans le kit Agilent qui permettent le séquençage en paire (« *paired-end* » en anglais), c'est-à-dire débutant aux deux extrémités des brins d'ADN séquencés. Ce type de séquençage permet de doubler le nombre de séquences obtenues par rapport à un séquençage simple, mais aussi un alignement plus précis des lectures, notamment dans les régions d'ADN répété qui entraînent des difficultés de séquençage.

3.8 Préparation de la « *flow cell* » pour le séquençage

En pratique, une trousse NextSeq 500/550 à usage unique est nécessaire pour effectuer une analyse de séquençage sur le NextSeq 550. Chaque trousse comporte une « *flow cell* » (que l'on pourrait traduire par station microfluidique en français) ainsi que les réactifs nécessaires pour une analyse de séquençage. La *flow cell*, la cartouche de réactifs et la cartouche de tampon utilisent une identification par radiofréquence (RFID) pour un suivi précis des consommables et pour des questions de compatibilité.

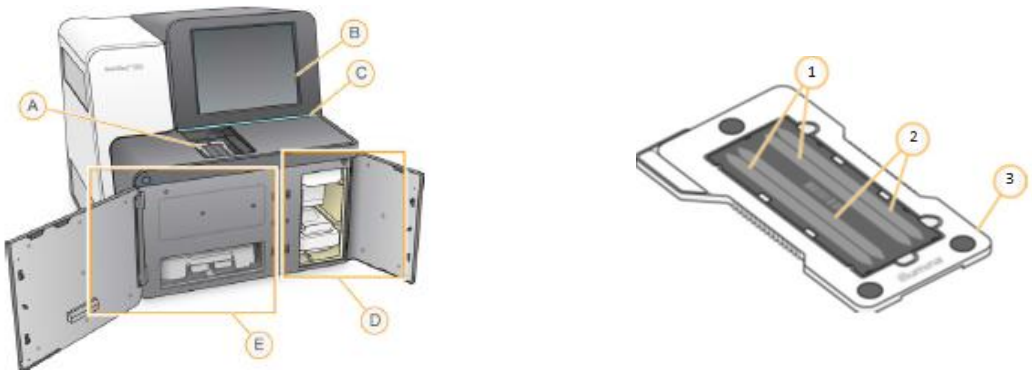


Figure 5. Système NextSeq 550 et cartouche de *flow cell* enchâssée (*d'après des images Illumina*)

A : Compartiment d'imagerie : contient la flow cell pour le séquençage ou l'adaptateur de puce BeadChip pour le balayage.
B : Moniteur tactile : permet la configuration et le paramétrage sur l'instrument à l'aide de l'interface du logiciel de commande.
C : Barre d'état : indique si l'instrument est en cours de traitement (bleu), s'il nécessite une attention particulière (orange), s'il est prêt pour le séquençage (vert) ou si un lavage doit être effectué dans les 24 prochaines heures (jaune).
D : Compartiment du tampon : contient la cartouche de tampon et le réservoir de réactifs usagés.
E : Compartiment de réactifs : contient la cartouche de réactifs.
1 : Paire de lignes A : lignes 1 et 3 ;
2 : Paire de lignes B : lignes 2 et 4 ;
3 : Châssis de la cartouche de flow cell.

La *flow cell* est un substrat de verre qui sert de support à la génération des amplicons et à la réaction de séquençage. Elle est enchâssée dans une cartouche et contient quatre lignes qui sont analysées par paires. Les lignes 1 et 3 (paire de lignes A) sont analysées simultanément. Les lignes 2 et 4 (paire de lignes B) sont analysées lorsque l'analyse de la paire de lignes A est terminée.

Avant le séquençage proprement dit, le chargement des librairies sur la *flow cell* doit être réalisé (voir schéma Figure 6) : l'ADN est fragmenté et des adaptateurs sont liés à ses extrémités de façon aléatoire pour permettre l'attachement de l'ADN à la *flow cell*. L'ajout de nucléotides non marqués et d'enzymes permet d'initier l'amplification « en pont » du brin attaché. Un brin complémentaire libre est alors synthétisé, formant des fragments double-brins. Ceux-ci seront dénaturés à nouveau pour pouvoir commencer une nouvelle amplification en pont et ainsi de suite, pour permettre *in fine* la formation de « *clusters* », c'est-à-dire de groupes de séquences identiques (ce sont des « clones ») sur la *flow cell*.

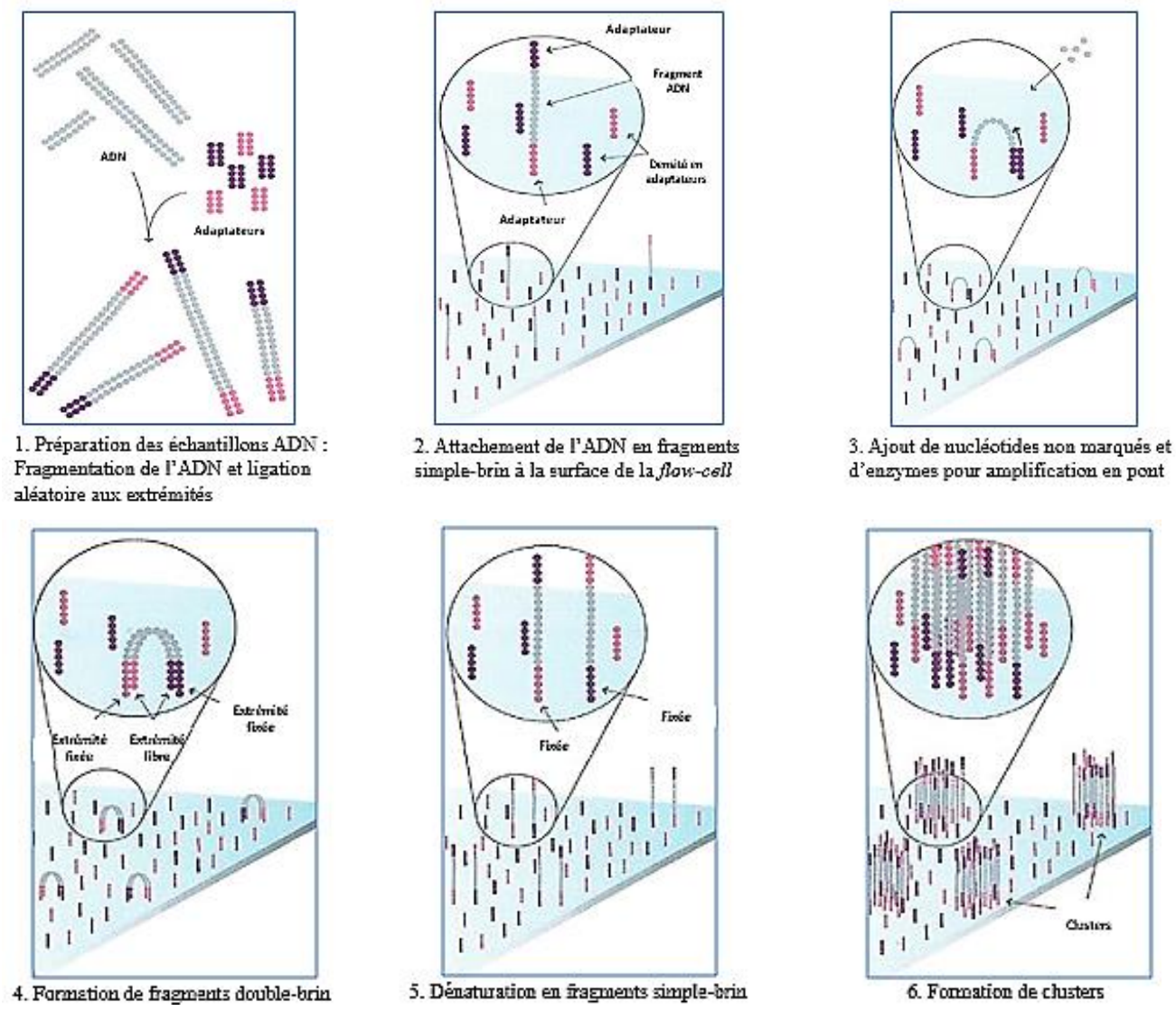


Figure 6. Etape 2 : Préparation des librairies pour le séquençage ***(d'après des images Illumina)***

3.9 Réaction de séquençage

Le séquençage est paramétré grâce au NextSeq Control Software, le logiciel de commande. Le logiciel Real-Time Analysis effectue l'analyse des images et la définition des bases lors de l'analyse. Sa durée varie selon le nombre de cycles réalisés. Pour notre expérimentation, elle comporte 150 cycles de séquençage et dure environ 30 heures.

Le séquençage se divise en plusieurs cycles chimiques (Figure 7) ; pour initier le premier cycle, on ajoute dans la *flow cell* des amorces nucléotidiques, de l'ADN polymérase et des nucléotides marqués à l'aide de fluorochromes qui arrêteront temporairement la réaction de manière réversible. Après une excitation laser, on peut capturer une image numérique de la fluorescence émise par chaque *cluster*, et enregistrer l'identité de la première base de chacun d'entre eux. Les autres cycles chimiques se déroulent sur le même mode.

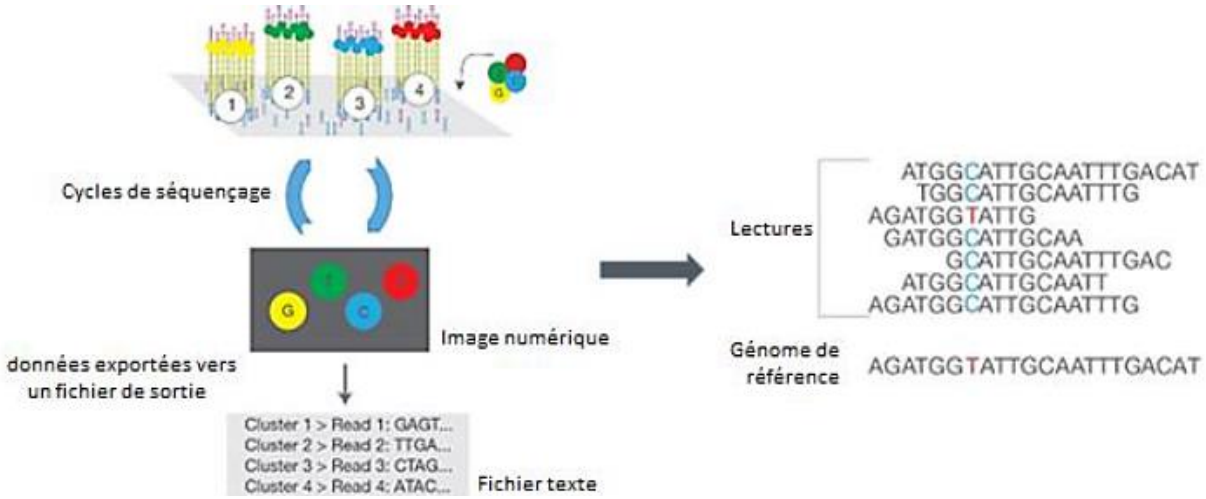


Figure 7. Etape 3 : Réaction de séquençage (*d'après des images Illumina*)

La quantité d'ADN chargée est critique pour la qualité du séquençage ; une *flow cell* trop ou trop peu chargée ne permettra pas un séquençage correct de l'exome. En effet, trop peu de matériel issu des librairies ne permet pas la bonne constitution des *clusters*. A l'inverse, une *flow cell* trop chargée comportera beaucoup de *clusters*, qui pourraient entrer en conflit. Ces deux situations peuvent entraîner des difficultés dans l'identification des bases nucléotidiques lors du séquençage.

4. Etapes bioinformatiques de l'analyse NGS

4.1 Recueil des données

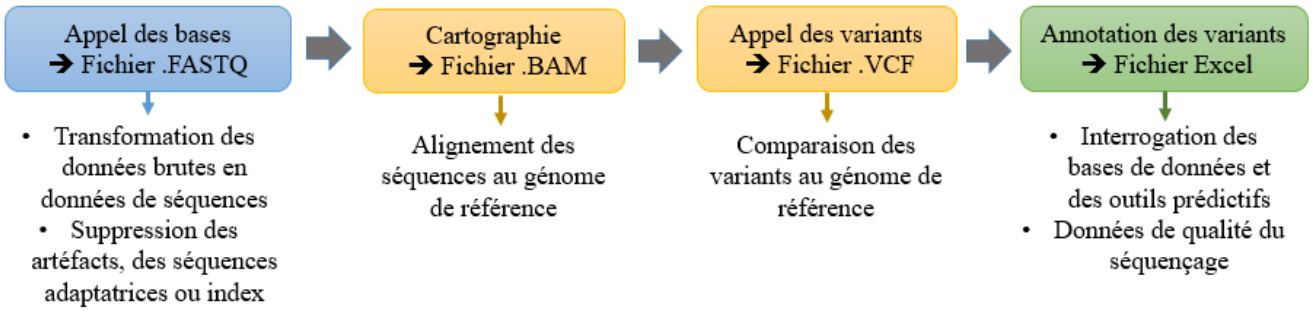


Figure 8. Etapes simplifiées de l'analyse NGS

La réalisation des librairies est effectuée en deux jours en moyenne. Le séquençage proprement dit (sur la plateforme NextSeq 550 du CHU de Poitiers) dure 30 heures. Les données brutes obtenues à l'issue de ce séquençage sont converties en environ 12 à 14 heures en données de séquence d'ADN durant l'étape d'appel des bases (« *base calling* » en anglais).

Pendant la progression de l'analyse, le logiciel de commande du séquenceur transfère automatiquement les fichiers de définition des bases vers le serveur d'analyse, où ils seront convertis en format FASTQ. Ce format de fichier texte permet de stocker à la fois des séquences biologiques (uniquement des séquences d'acides nucléiques) et les scores de qualité associés **;** c'est l'analyse primaire, puis il y a une élimination des séquences des amorces de séquençage dans les lectures obtenues. Les différentes séquences, appelées encore lectures ou « *reads* » sont alignées sur un génome de référence grâce à l'utilisation en cascade de plusieurs logiciels bioinformatiques (par exemple SolexaQA, les outils SAM (*Sequence Alignment/Map*), MarkDuplicates), pour mettre en évidence d'éventuelles variations.

A l'issue de ces deux étapes d'alignement et de comparaison, qui durent respectivement 10 et 24 heures, on obtient deux types de fichiers :

- les fichiers .BAM (*Binary Alignment Map*), issus des données de cartographie (« *mapping* »), qui permettent de calculer la profondeur de lecture ainsi que la couverture des régions d'intérêt (définitions ci-après paragraphe 4.3)
- le fichier .VCF (*Variant Call Format*), qui durant l'étape d'appel des variants (« *variant calling* ») met en évidence toutes les différences de la séquence du patient avec le génome de référence, soit l'ensemble des variants de chaque patient.

4.2 Qualité du séquençage et annotation des variants

L'analyse finale des données, correspondant à l'étape de l'interprétation des variants, s'avère cruciale et délicate. Elle est effectuée en collaboration avec un Bio-informaticien, en particulier toute la partie concernant le tri et la priorisation des variants nucléotidiques issus du fichier destiné à l'utilisateur final.

L'annotation des variants à l'aide de nombreux outils informatiques est nécessaire à leur interprétation. Le but de cette étape est d'obtenir un maximum d'informations concernant chaque variant par l'intégration des données issues des bases de données publiques telles que GnomAD (http://gnomad.broadinstitute.org/), EVS (http://evs.gs.washington.edu/EVS/) et dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/), pour apprécier la fréquence de ce variant dans diverses populations (contrôles, malades, etc…) et sa potentielle pathogénicité (établie sur des scores de prédiction multiples : conservation, propriétés physico-chimiques, données structurales, séquences régulatrices, etc…).

L'annotation comprend des indications sur la région dans laquelle le variant se situe (exonique, intronique, sur un site d'épissage…), sur le gène (ou séquence d'intérêt) concerné, sur le changement résultant dans le codage de l'information (variation faux-sens, non-sens, synonyme ou décalage du cadre de lecture…) mais aussi sur les données relatives à la qualité du séquençage et de la profondeur de lecture. Cette étape d'annotation dure environ 10 heures, grâce à la puissance de notre serveur informatique (128 CPU pour « *Central Processing Unit* », c'est la puissance de calcul du processeur, 1 Téraoctet (To) de RAM pour « *Random Access Memory* », soit la mémoire vive, et 40 Téraoctets de stockage).

Toutes ces données d'annotations sont indiquées pour chaque variant contenu dans le fichier .VCF, qui est produit pour le séquençage de l'exome d'un patient donné (contenant entre 40 000 et 50 000 variants). Ce sont ces données qu'analyse le généticien biologiste pour déterminer la pathogénicité présumée d'un variant nucléotidique, en tentant d'établir une corrélation phénotype/génotype avec la clinique présentée par le patient.

4.3 Notions de couverture et de profondeur de lecture

Avant l'étape finale d'analyse, nous portons une attention particulière à la qualité du séquençage, c'est-à-dire à la couverture de l'exome et à la profondeur de lecture.

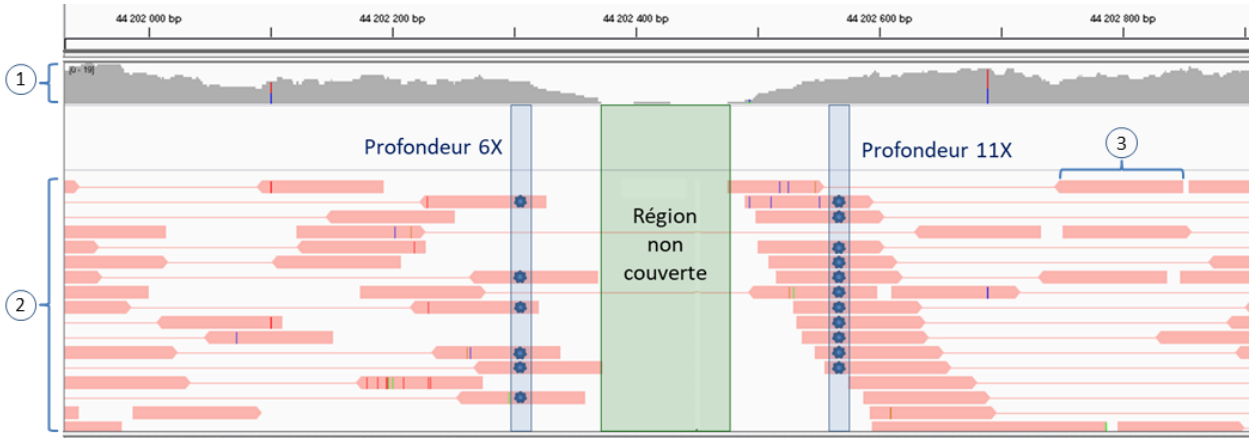


Figure 9. Illustration des notions de couverture et de profondeur de lecture
1 : Représentation du nombre de lectures de chaque base de la séquence
2 : Ensemble des séquences obtenues par séquençage haut débit
3 : Représentation d'une lecture sur le fichier .BAM

La couverture de l'exome correspond au nombre moyen de lectures qui s'alignent ou « couvrent » les bases (connues) de la séquence de référence. Une région non couverte est une région dont la profondeur de lecture est insuffisante et donc une région où l'identification d'un variant est impossible. La profondeur de lecture se rapporte au nombre de fois où une base est « lue » dans les lectures obtenues. On parle de base lue « N fois » ou « Nx ». En dessous de « 20x », les variants sont considérés « faibles » du point de vue technique (c'est-à-dire présents dans une minorité des lectures) et ne doivent pas être pris en compte dans l'analyse biologique.

Dans notre projet, nous recherchons une couverture et une profondeur optimales ; par exemple, nous souhaitons obtenir une profondeur de lecture minimale de 20x pour 95% des séquences de chaque exome. Idéalement, dans un contexte diagnostic, il faudrait atteindre une profondeur minimum de 20x pour une couverture de 100%, selon les recommandations de l'*European Society of Human Genetics* (ESHG), (Matthijs G *et al*., 2016).

5. Attribution d'un score à chaque variant

Actuellement, au Laboratoire de Génétique Moléculaire, les variants annotés sont visualisés à l'aide d'un tableur de type Excel (Microsoft®), ce qui permet de les filtrer selon diverses conditions (fréquence dans la population générale, implication dans une pathologie répertoriée dans OMIM, mode de transmission si connu…). Mais cette méthode est longue (du fait du nombre de variants générés par l'analyse), fastidieuse et potentiellement délétère, car en fonction du filtre appliqué, elle peut entraîner la mise à l'écart de variants qui peuvent constituer de bons candidats pour expliquer la clinique présentée.

Dans ce contexte de délais importants d'analyse et de remise des résultats biologiques aux cliniciens et *a fortiori* aux patients, le développement d'un système plus rapide pour l'analyse

biologique est crucial. Pour ce faire, nous proposons une nouvelle annotation à chaque variant, « résumant » ses caractéristiques propres, sous la forme d'un score.

L'application nommée « GenSCor » (Q. Riché-Piotaix *et al*, communication personnelle) nous permet de construire les règles pour attribuer un métascore à chaque variant. A partir des données des différentes annotations, nous les convertissons en valeurs numériques, pondérées selon notre choix. En pratique, ce module nous permet de tester de manière autonome sur chaque variant des hypothèses de façon indépendante (règle simple) ou interdépendante (règles multiples). Ce logiciel permet un gain en efficience et en efficacité par rapport à une analyse utilisant les filtres d'un tableur classique.

Pour établir un ensemble de règles fiable pour trier les données d'exome à venir, nous avons repris les analyses d'exomes antérieures où un variant a déjà été identifié comme pathogène au vu des différentes annotations et considéré comme responsable de la pathologie présentée, car on peut établir avec certitude une corrélation génotype/phénotype pour le patient concerné.

Au vu du nombre de variants à trier dans un seul exome (40 000 à 50 000), il apparaît nécessaire d'établir plusieurs ensembles de règles, selon les 3 hypothèses majoritaires de transmission : un premier pour faire ressortir des variants candidats selon un mode de transmission autosomique dominant, un deuxième pour un mode de transmission lié à l'X et un troisième pour un mode de transmission autosomique récessif.

Nous disposons de 21 exomes considérés comme « positifs » : 14 contiennent un variant pathogène identifié comme de transmission autosomique dominante (AD), 5 avec un variant pathogène de transmission liée à l'X, et 2 avec un variant pathogène considéré comme de transmission autosomique récessif (AR). La construction de chacun des trois ensembles de règles correspondant au mode de transmission a été effectuée grâce à l'utilisation des données de ces exomes « contrôles ».

6. Etude de la transmission allélique des variants des gènes candidats identifiés

Pour confirmer la présence des variants dans les gènes candidats sélectionnés et étudier la transmission allélique au niveau parental, nous utilisons des amorces d'oligonucléotides spécifiques des régions concernées pour effectuer un séquençage selon Sanger. Ces amorces sont élaborées à l'aide du logiciel OLIGO Primer Analysis 2.0.

IV. Résultats / Discussion

1. Point sur la couverture et la profondeur de lecture

Sur l'ensemble des exomes que nous avons analysés, la profondeur de lecture moyenne est de 111x, avec une profondeur de lecture minimale de 20x pour 94,5% des séquences de chaque exome, soit 0,5% de moins que la limite inférieure minimale que nous nous étions fixés. Cela peut s'expliquer par la quantité d'ADN utilisée pour la technique de l'exome.

En effet, avec l'utilisation de notre kit actuel, cette quantité conditionne la qualité de la couverture de l'exome analysé. Par exemple, la concentration d'ADN à l'initiation de la technique doit être comprise entre 22 et 28 ng/µL, soit un intervalle très étroit. Il est donc nécessaire de vérifier à différentes reprises les concentrations avant et pendant l'expérimentation, notamment à l'aide de la TapeStation® et du Qubit™.

Ce sont toujours les mêmes zones qui restent non ou mal couvertes dans l'analyse de l'exome. Ce sont en général des régions riches en nucléotides G et C ou des régions de séquences répétées, provoquant des difficultés de séquençage. Ce défaut de couverture provient aussi du kit que nous utilisons pour la capture des exons.

Par exemple, il y a 85 exons de gènes qui ne sont pas entièrement couverts par l'analyse, dont deux font partie des régions codantes pour des gènes répertoriés dans la base de données OMIM, *MFRP* (impliqué dans une forme de microphtalmie de transmission autosomique récessive) et *NLRP3* (identifié dans plusieurs syndromes cliniques dont la présentation comprend entre autres des symptômes inflammatoires chroniques, des signes cutanés, ainsi qu'une surdité). Ces régions mal couvertes ne permettent donc pas l'identification d'un variant potentiellement pathogène.

2. Exemple de résultat obtenu après une analyse d'exome

Un exemple extrait de fichier .VCF recueilli après analyse bio-informatique, annotation et attribution d'un score aux variants, sous forme d'un tableur Excel (Microsoft®), est disponible en Annexe 2. Chaque ligne correspond à un seul variant ; le tableur en compte 40 000 à 50 000 pour chaque analyse. Il est constitué actuellement de 124 colonnes, chacune correspondant à une annotation du variant, provenant d'une base de données particulière.

Le score attribué en fonction de l'ensemble de règles choisi est visible dans la 1ere colonne. Il y a 29 colonnes qui correspondent aux informations sur l'identité du variant (localisation précise, gène concerné, nomenclature). Six colonnes reportent les annotations concernant la fréquence à laquelle ce variant est répertorié, dans notre propre base de données et dans les bases internationales ; 10 autres

sont des indicateurs contrôles de la qualité du séquençage (couverture, profondeur de lecture…) ; 4 concernent l'hypothèse de transmission associée au variant, issue de la base de données OMIM ; 21 colonnes recensent les annotations de pathogénicité potentielle répertoriée dans les bases de données. Les 53 colonnes restantes contiennent des informations sur les données obtenues avec l'utilisation de scores d'intégration et d'outils informatiques de prédiction de pathogénicité. C'est l'ensemble de ce tableur qui sert de base de travail au biologiste pour l'interprétation des variants.

3. Résultats de l'étude rétrospective

3.1 Elaboration des scores à l'aide des exomes contrôles

L'élaboration des scores repose sur un certain nombre de règles indépendantes, introduites dans GenSCor et pondérées positivement ou négativement par le biologiste. Un exemple de visualisation des règles dans l'application GenSCor est disponible en Annexe 3. Les principales règles et leur pondération sont résumées dans le Tableau 5.

Critère de la règle	Caractéristique du variant	Pondération des colonnes
Localisation du variant	région non codante (hors région d'épissage)	- - -
Nature du variant	non-sens, stop prématuré, décalage du cadre de lecture, site canonique d'épissage	+ + +
	faux-sens	+ +
	Synonyme	+
Fréquence dans la population	variant rare (fréquence < 1/10 000)	+ +
	variant fréquent (fréquence > 3/10 000)	- - -
Qualité du séquençage	qualité faible	- - -
Hypothèse de transmission	fonction du score choisi (AD, AR ou liée à l'X)	+
Classification si variant connu dans les bases de données	pathogène ou probablement pathogène	+
	bénin ou probablement bénin	-
	rôle connu dans la DI	+

+ : pondération positive ; - : pondération négative
1 indicateur : faible pondération ; 2 indicateurs : pondération intermédiaire ; 3 indicateurs : pondération importante

Tableau 5. Principales règles établies pour l'élaboration des scores

L'analyse de l'exome comprend l'étude des régions codantes du génome ainsi que des régions flanquantes des exons, qui sont des régions introniques notamment utiles à l'épissage. L'interprétation des autres régions est pour le moment trop délicate à effectuer et les variants dans ces régions présentent donc peu d'intérêt. Un variant fréquemment rapporté dans la population générale est par définition identifié comme un polymorphisme, nous nous intéresserons donc exclusivement aux variations rares, ce qui explique la pondération négative importante des variants considérés comme fréquents.

Plusieurs bases de données concernant les pathologies, telle ClinVar (http://www.ncbi.nlm.nih.gov/clinvar), répertorient les variations nucléotidiques et leur corrélation phénotypique dans les différentes publications et les classent selon les critères de l'*American College of Medical Genetics* en vigueur (Richards *et al*., 2015).

Un variant recensé comme probablement bénin ou bénin présente beaucoup moins d'intérêt qu'un variant identifié comme pathogène dans plusieurs publications, surtout avec une implication clairement identifiée dans la DI, c'est pourquoi ce dernier dispose d'une pondération positive. De même, l'ACMG introduit des critères de pathogénicité selon la nature de la variation : un variant engendrant une perte de fonction du gène est souvent très délétère alors que la preuve de la pathogénicité d'un variant synonyme est beaucoup plus difficile à établir.

Nous avons établi trois scores en fonction de l'hypothèse de transmission des variants : un score AD pour une transmission autosomique dominante, un score X pour une transmission liée à l'X, et un score AR pour une transmission autosomique récessive. Les colonnes se rapportant à l'hypothèse de transmission allélique sont pondérées différemment en fonction du score choisi : par exemple pour le score AD, nous recherchons un variant avec une hypothèse de transmission AD, c'est donc une information relative à cette hypothèse de transmission qui sera pondérée positivement dans les colonnes se rapportant à la transmission allélique.

Dans le score X, les variants situés sur le chromosome X ont bénéficié d'une pondération très importante, les autres n'ayant pas d'intérêt particulier dans ce type de transmission. Dans le score AR, ce sont les variants présentant un haplotype homozygote qui ont bénéficié d'une forte pondération positive (car dans la plupart des cas, ce type de transmission s'observe dans les familles consanguines). Pour ce score également, nous avons pondéré d'autres colonnes du tableur contenant des données sur la prédiction de la pathogénicité du variant ; ce sont en fait des scores d'intégration de prédiction fournies par des bases de données spécifiques, qui permettent de mettre en évidence les variations prédites délétères.

3.2 Evaluation du score AD

N° patient	Nom du gène	Nomenclature (transcrit de référence)	Haplotype	Type de score	Score du variant	Rang du variant	Interprétation clinico-biologique
275	*DYNC1H1*	c.388T>C	Hétérozygote	AD	450	3	classe 4
298	*CACNA1A*	c.1762C>T	Hétérozygote	AD	450	3	classe 4
326	*AUTS2*	c.901C>T	Hétérozygote	AD	600	1	classe 5
327	*CTCF*	c.782-2A>G	Hétérozygote	AD	500	3	classe 5
333	*TRRAP*	c.3127G>A	Hétérozygote	AD	350	10	classe 3
348	*DLG4*	c.1978C>T	Hétérozygote	AD	500	2	classe 5
366	*KIF11*	c.2548-1G>A	Hétérozygote	AD	600	1	classe 5
383	*CDK13*	c.2525A>G	Hétérozygote	AD	550	1	classe 5
385	*SLC32A1*	c.965T>G	Hétérozygote	AD	350	17	classe 3
386	*PTPN11*	c.853T>C	Hétérozygote	AD	650	1	classe 5
387	*EP300*	c.5571_5578delACCAACTG	Hétérozygote	AD	500	1	classe 5
388	*SETD2*	c.4043delA	Hétérozygote	AD	350	7	classe 5
391	*CHD3*	c.645C>T	Hétérozygote	AD	500	2	classe 5
392	*SETD1*	c.4876C>T	Hétérozygote	AD	450	3	classe 4

Tableau 6. Variants identifiés comme « pathogènes » de transmission AD dans l'étude rétrospective

Ce tableau résume les caractéristiques de score des variants désignés comme responsables de la pathologie à l'issue d'une concertation clinico-biologique, c'est-à-dire interprétés comme des variants de classe 4 et classe 5, soit respectivement « probablement pathogènes » et « pathogènes » selon la classification de l'ACMG.

Parmi ces variants, deux sont de classe 3 (en rose), soit identifiés comme des variants de signification indéterminée (VUS, *Variant of Unknown Significance* en anglais). Ceux-ci sont des variants pour lesquels la présomption clinique de pathogénicité est très forte en s'appuyant sur la littérature. Les variants ont au minimum un score de 350, pour un rang maximum à 17.

Depuis l'analyse de cet exome, le variant *de novo* dans le gène *TRRAP* a fait l'objet d'une étude multicentrique publiée en février 2019 (Cogné *et al*., 2019). Dans cette étude, 17 variants de ce gène impliqué dans le complexe d'histone acétyltransférase ont été étudiés, avec l'hypothèse d'une transmission autosomique dominante, et reconnus pathogènes dans le cadre d'une association avec un phénotype incluant DI syndromique et autisme chez les patients. La classification de ce variant devrait donc évoluer vers une classe supérieure.

L'annotation n'étant toutefois pas mise à jour qu'au regard de cette seule publication, des informations manquent dans les bases de données pour l'annotation informatique du fichier .VCF à

interpréter à l'issue de l'analyse. Parmi ces informations manquantes, le mode de transmission n'est par exemple pas précisé ; la pondération de cette colonne « mode de transmission » pour ce variant n'est donc pas effective, ce qui explique le score moindre et son rang éloigné par rapport aux scores et aux rangs des variants de classe 4 et 5.

Le gène *SLC32A1* code pour une protéine qui est un co-transporteur impliqué dans le transport vésiculaire du GABA (*gamma aminobutyric acid* soit acide γ-aminobutyrique en français), intervenant ainsi dans l'homéostasie corticale (Fattorini *et al*., 2015). La variation identifiée dans ce gène est une variation faux-sens, survenue *de novo* et non répertoriée dans les bases de données. Elle affecte un acide aminé au niveau de la séquence codante pour un domaine transmembranaire de la protéine, très conservée dans l'évolution. De plus, elle est prédite très délétère par plusieurs logiciels de prédiction *in silico*, même si ce gène n'est pas encore décrit en pathologie humaine.

Le motif de la réalisation de l'analyse chez cette patiente est une encéphalopathie convulsivante et, compte-tenu des informations recueillies, la variation a été retenue comme responsable de la pathologie par corrélation phénotype/génotype en réunion clinico-biologique, avec un caractère de transmission autosomique dominant.

Compte tenu des caractéristiques de ce variant relevant de la recherche, il est défini comme de classe 3. L'annotation de certaines des particularités de ce variant, comme la transmission AD, n'étant pas disponible encore une fois dans les bases de données, elle n'est pas présente sur le fichier .VCF et n'est pas pondérée par le score établi pour les variants AD. Ceci explique à nouveau le score relativement bas et le rang éloigné par rapport aux variants de classe 4 et 5.

3.3 Evaluation du score X

N° patient	Nom du gène	Nomenclature (transcrit de référence)	Haplotype	Type de score	Score du variant	Rang du variant	Interprétation clinico-biologique
277	*MECP2*	c.842delG	Hémizygote	X	300	1	classe 5
307	*FRMPD4*	c.3965-2A>C	Hémizygote	X	300	1	classe 4
360	*CACNA1F*	c.3360_3361delCA	Hémizygote	X	300	2	classe 5
278	*HDAC8*	c.473C>A	Hétérozygote	X	250	3	classe 4
381	*F8*	c.2050G>T	Hémizygote	X	250	2	classe 4

Tableau 7. Variants identifiés comme « pathogènes » de transmission liée à l'X dans l'étude rétrospective

Pour un total de 5 variants de classe 4 et 5 identifiés de transmission liée à l'X, hémizygotes ou hétérozygotes, les scores sont de 250 ou 300, avec un rang élevé de 1 à 3, c'est-à-dire que les variants sont retrouvés dans les premières lignes du tableur de variants annotés pour chacun de ces patients.

3.4 Evaluation du score AR

N° patient	Nom du gène	Nomenclature (transcrit de référence)	Haplotype	Type de score	Score du variant	Rang du variant	Interprétation clinico-biologique
337	*TAF8*	c.781-1G>A	Homozygote	AR	550	7	classe 5
349	*QARS*	c.134G>T	Hétérozygote	AR	650	1	classe 5
349	*QARS*	c.727T>G	Hétérozygote	AR	350	97	classe 4

Tableau 8. Variants identifiés comme « pathogènes » de transmission AR dans l'étude rétrospective

Le patient 33717 présente un variant à l'état homozygote, hérité des deux allèles parentaux, sur le gène *TAF8*. Cette variation qui détruit un site accepteur d'épissage n'est pas répertoriée dans les bases de données mais est décrite dans une étude très récente (El-Saafin *et al*., 2018) chez un patient atteint de DI dont la production de la protéine TAF8 est nulle. Notre patient présentant lui-même une microcéphalie avec retard psychomoteur, la corrélation phénotype/génotype établie a permis de classer ce variant comme de classe 5.

Le patient 34926 présente deux variants du gène *QARS* à l'état hétérozygote, soit hétérozygote composite. Le variant c.134G>T hérité de sa mère, est répertorié dans la base de données ClinVar comme « pathogène » et a été rapporté en 2014 (Zhang *et al*., 2014) chez deux frères présentant une encéphalopathie épileptique pharmaco-résistante, une microcéphalie évolutive et une atrophie cérébrale. Il est identifié comme variant de classe 5. Le second variant c.727T>G hérité du père n'est pas répertorié dans les banques de données mais il est prédit délétère d'après plusieurs logiciels de prédiction *in silico*.

Comme il existe une très bonne corrélation phénotype/génotype avec notre patient (microcéphalie évolutive, encéphalopathie épileptique pharmaco-résistante associée à une simplification gyrale et corps calleux atrophié), il est indiqué de le répertorier en classe 4. Néanmoins il manque des données d'annotation pour ce variant dans le fichier .VCF du fait qu'il soit inconnu des bases de données, le score de 350 restant faible par rapport aux autres variants et son rang demeure très éloigné (97).

3.5 Effet « seuil » du score

Dans l'étude rétrospective, nous avons remarqué un effet « seuil » par rapport au score établi. En effet, les variants de classe 4 ou 5 ont un score « seuil » de 350 avec le score AD, de 250 avec le score X (soit 50% du score maximum que l'on peut atteindre, respectivement de 700 et de 500 pour

chaque ensemble de règles). Pour le score AR, ce seuil n'est pas identifiable, puisque ce score a été construit en fonction de 2 analyses seulement.

Cependant, cette notion de seuil est utile pour l'interprétation des variants, puisque si ces scores sont fiables, il n'est alors plus nécessaire de considérer les variants inférieurs au seuil comme potentiellement pathogènes dans un cadre diagnostic. Il en résulte un gain de temps non négligeable pour l'interprétation biologique, avec la production d'un tableur de variants déjà trié en fonction du score.

4. Résultats de l'étude prospective

4.1 Variants pathogènes et variants « recherche » retenus

Pour l'étude prospective, nous avons analysé 78 exomes. Nous avons exclu 16 exomes de patients car ne présentant pas de DI et/ou de retard des acquisitions (RA). Parmi les 62 exomes analysés, nous avons effectué toutes les étapes de la technique pour 48 d'entre eux, en incluant 5 patients recrutés pour le PHRC « Etude clinique, neuropsychologique et moléculaire du syndrome CHARGE ». Au cours de ces analyses, 10 variants (Tableau 9) ont été identifiés lors d'une concertation clinico-biologique comme pathogènes d'emblée, soit un rendement diagnostic de 16%. Parmi ceux-ci, 7 variants sont de transmission AD (avec un score variant de 250 à 700), 2 de transmission liée à l'X (scores de 350 et 450) et 1 de transmission AR (score de 650). Les variants candidats dans le cadre de la recherche ont été également retenus lors de cette réunion.

N° patient	Sexe	Gène	Haplotype	Type de score	Score du variant	Rang du variant	Nomenclature (transcrit de référence)	Variation protéique	Type de variation
276	F	*GNB1*	Hétéro.	AD	550	1	c.227A>G	p.Asp76Gly	faux-sens
353	F	*BCL11A*	Hétéro.	AD	500	2	c.599_602delAAAG	p.Glu200fs	DCL
354	M	*NTRK2*	Hétéro.	AD	250	61	c.2356C>T	p.Arg786Ter	non-sens
373	M	*PURA*	Hétéro.	AD	500	2	c.648delC	p.Glu217fs	DCL
382	F	*PTPN11*	Hétéro.	AD	550	1	c.182A>G	p.Asp61Gly	faux-sens
389	F	*AUTS2*	Hétéro.	AD	600	1	c.376C>T	p.Arg126Ter	non-sens
413	F	*SIN3A*	Hétéro.	AD	700	1	c.1675C>T	p.Arg559Ter	non-sens
374	M	*DYM*	Homo.	AR	650	2	c.1938_1942delTGTCT	p.Val647fs	DCL
376	F	*HDAC8*	Hétéro.	X	350	1	c.908G>A	p.Gly303Glu	faux-sens
404	M	*NAA10*	Hémi.	X	450	5	c.206A>C	p.His69Pro	faux-sens

Hétéro. : hétérozygote ; Homo. : homozygote ; Hémi. : hémizygote ; AD : autosomique dominant ; AR : autosomique récessif ; X : transmission liée à l'X ; DCL : décalage du cadre de lecture

Tableau 9. Récapitulatif des variants identifiés comme « pathogènes » dans l'étude prospective

Un seul variant de transmission AD présente un score de 250, soit inférieur au seuil observé. Il s'agit d'un variant non-sens dans le gène *NTRK2* (en orange). Dans la base de données OMIM, des corrélations phénotype/génotype ont été établies entre des variants retrouvés dans ce gène et des phénotypes particuliers, avec un mode de transmission AD. Pourtant, cette information n'est pas rapportée dans les colonnes du tableur dédiées au mode de transmission. Il s'agit d'une erreur survenue dans l'annotation du variant à partir de bases de données inexactes, ce qui explique le score relativement faible obtenu pour ce variant avec l'ensemble de règles AD. Une mise à jour des bases de référence permettrait de faire remonter le score de ce variant à la valeur seuil observée.

Au total, 15 variants « recherche » candidats ont été retenus et leur transmission allélique a été étudiée au moins partiellement pour 12 d'entre eux. Presque tous ces variants ont une hypothèse de transmission autosomique dominante, seul un variant avec une hypothèse de transmission récessive a été mis en évidence avec les scores établis. L'étude de ces variants chez les parents a permis de montrer que bon nombre étaient bénins car transmis par un parent « non atteint », qui ne présente pas de DI.

Les explorations pour ces variants hérités ont été abandonnées. Nous avons en revanche identifié 4 variants qui présentent un intérêt pour des explorations futures et qui sont répertoriés dans le tableau 10.

N° patient	Présentation clinique	Gène	Haplo.	Type de score	Score du variant	Rang du variant	Nomenclature (transcrit de référence)	Variation protéique	Type de variation	Transmission allélique
200	RA + troubles du comportement	***USP19***	Hétéro.	AD	350	9	c.2726G>A	p.Cys909Tyr	Faux-sens	De novo
343	DI syndromique + épilepsie	***NCKAP1***	Hétéro.	AD	500	2	c.2410C>T	p.Arg804Ter	Non-sens	Mère exclue
358	DI syndromique	***FKBP4***	Hétéro.	AD	350	8	c.127G>A	p.Gly43Ser	Faux-sens	Père exclu
270	PHRC CHARGE	***ESRP1***	Homo.	AR	350	62	c.1552G>T	p.Val518Phe	Faux-sens	NA

DI : déficience intellectuelle ; RA : retard des acquisitions ; Haplo. : haplotype ; Hétéro. : hétérozygote ; Homo. : homozygote
AD : autosomique dominant ; AR : autosomique récessif ; NA : non analysée

Tableau 10. Récapitulatif des variants « recherche » candidats

4.2 Variation faux-sens dans le gène *USP19*

La patiente 20001 présente à l'état hétérozygote une variation faux-sens p.Cys909Tyr dans l'exon 19 du gène *USP19* (NM_001199160). Le score AD de cette variation est de 350, le rang étant à 9. La protéine USP19 (*Ubiquitin-specific protease 19*) encodée par ce gène appartient à la famille

des enzymes protéases ubiquitine-spécifiques, qui clivent l'ubiquitine des substrats protéiques. Cette protéine se localise dans le noyau cellulaire et est exprimée dans tous les tissus, particulièrement au niveau du foie et du thymus. Elle joue un rôle dans la signalisation UPR (*Unfolded protein response*) par son implication dans le système ERAD (*Endoplasmic reticulum associated degradation*) (Hassink *et al.*, 2009). Cette variation n'est pas répertoriée dans les bases de données et elle est survenue *de novo* dans une région codante pour le domaine protéase spécifique de l'ubiquitine et pour un domaine doigt de zinc, régions protéiques qui sont très conservées dans l'évolution.

Nous avons déposé cette variation dans GeneMatcher, un outil développé par le Centre Baylor-Hopkins (Sobreira *et al*., 2015) accessible gratuitement en ligne et qui permet d'identifier d'autres individus présentant un phénotype rare qui ont eux aussi un variant de signification clinique inconnue dans le gène candidat. La portée internationale de cette plateforme nous a permis de retrouver deux cas index avec un variant rapporté dans ce gène. Nous avons contacté les rapporteurs pour avoir plus de renseignements concernant la variation et le phénotype présentés par leurs patients.

Nous avons obtenu une réponse du Dr Eldomery, de l'Université de Médecine Baylor à Houston. Chez un patient présentant une encéphalopathie épileptique, un retard de développement et une hypotonie, son équipe a identifié grâce à la technique de l'exome deux variations faux-sens (hétérozygotie composite) qui pourraient être impliquées dans le phénotype présenté par ce patient (Eldomery *et al.*, 2017). Notre patient présente lui un retard des acquisitions et des troubles du comportement.

De plus, un autre gène de la même famille, *USP34*, est exprimé dans le cerveau et interviendrait dans des processus neuro-développementaux comme la prolifération des neurones et des cellules gliales et la migration neuronale, par la voie de signalisation Wnt/β-caténine (Lui *et al.*, 2011). Il s'agit d'un gène candidat pour expliquer la DI dans le syndrome microdélétionnel 2p15-2p16 (Fannemel *et al*., 2014; Lévy *et al*., 2017).

Cependant, nous ne disposons pas à l'heure actuelle de données issues d'études fonctionnelles ni de suffisamment de patients présentant un variant dans la région du gène *USP19* qui pourraient nous confirmer l'hypothèse d'une transmission allélique autosomique dominante pour ce gène. Nous ne pouvons pas non plus exclure l'implication de ce gène dans la DI.

4.3 Variation non-sens dans le gène *NCKAP1*

Le patient 37327 présente à l'état hétérozygote une variation non-sens p.Arg804Ter dans l'exon 23 du gène *NCKAP1* (NM_205842). Le score AD de ce variant est de 500, pour un rang de 2. La protéine NCKAP1 (*NCK-associated protein 1*) est exprimée dans tous les tissus sauf les globules blancs périphériques, avec une forte expression dans le cerveau (prédominance dans l'amygdale et

l'hippocampe), le cœur et les muscles squelettiques. Elle est impliquée dans la survie neuronale. Elle fait partie du complexe WAVE, qui régule la réorganisation du filament d'actine par son interaction avec le complexe Arp2/3 (Steffen *et al*., 2004).

Ce gène est très intolérant à la perte de fonction. De plus, un variant tronquant dans l'exon 32 de ce gène a été identifié dans une grande famille présentant une forme modérée de DI non syndromique (Anazi *et al*., 2017), avec une transmission autosomique dominante. Toutefois, notre patient présente une DI modérée syndromique avec épilepsie associée et nous n'avons pu étudier la transmission allélique que partiellement. Une transmission maternelle a pu être exclue mais nous ne disposons que de très peu d'informations sur le père, dont le prélèvement reste indisponible.

Nous avons soumis ce gène dans GeneMatcher et nous avons établi un contact avec le Dr Hui Guo, du Département des Sciences du Génome de la faculté de médecine de Seattle. Il étudie actuellement 21 patients présentant des troubles du spectre autistique, une DI modérée, ou un RA, et dont l'analyse retrouve une variation non-sens, une délétion ou une inversion dans ce gène, héritée ou *de novo*. Notre patient pourrait donc correspondre par des caractéristiques communes au phénotype étudié (en excluant l'épilepsie). L'histoire familiale de notre patient fait aussi état d'une déficience intellectuelle chez son frère (associée à des troubles psychiatriques) et sa sœur (associée à des troubles du spectre autistique), dont nous ne disposons pas des prélèvements actuellement.

Nous avons recontacté la famille pour avoir des prélèvements et une collaboration scientifique nous a été proposée par le Dr Guo. De plus, il nous a précisé qu'une étude fonctionnelle sur un modèle de poisson-zèbre est en cours ; la délétion complète du gène *NCKAP1* serait létale selon les premiers résultats. Un modèle présentant une délétion hétérozygote de ce gène est à l'étude actuellement.

4.4 Variation faux-sens dans le gène *FKBP4*

Le patient 35877 présente à l'état hétérozygote la variation faux-sens p.Gly43Ser dans l'exon 2 du gène *FKBP4* (NM_002014). Le score AD de ce variant est de 350 avec un rang de 8. Ce gène est très exprimé au niveau du système nerveux central. La protéine FKBP4 (*T-cell FK506-binding protein*) se lie à l'immunosupresseur FK506 ou neuroimmunophiline. Elle possède des fonctions neuritotrophiques, neuroprotectrices et neurorégénératives (Hausch, 2015).

L'étude de la ségrégation familiale a pu exclure une transmission paternelle. Nous avons aussi étudié les deux frères de ce patient, qui présentent une DI eux aussi mais dont le phénotype n'est pas tout à fait identique, même s'ils partagent de caractéristiques similaires (DI légère syndromique, avec déformations thoraciques et micrognathie). Seul l'un des deux présente cette variation. Cette étude familiale demeure cependant incomplète, la mère étant décédée.

Nous avons soumis cette variation dans GeneMatcher et nous avons été contactés par le Pr Siddharth Banka de l'Université de Manchester, qui mène actuellement des études fonctionnelles sur ce gène. Les cas index étudiés par son équipe présentent une DI sévère et une ambiguïté génitale. Il semblerait que la pathologie dont ils sont atteints soit de transmission autosomique récessive, ces patients présentant une mutation bi-allélique perte de fonction dans ce gène.

Il n'y a donc pas de concordance phénotypique possible avec notre patient, qui présente une DI légère et quelques traits morphologiques particuliers, ainsi qu'une mutation faux-sens. Toutefois, les effets de ces variants dans le gène *FKBP4* sont encore méconnus aujourd'hui, on ne peut donc pas l'exclure en tant que gène candidat impliqué dans la DI.

4.5 Variation homozygote faux-sens dans le gène *ESRP1*

Le patient 27079 appartient à la cohorte recrutée pour le PHRC CHARGE. Il présente une variation faux-sens p.Val518Phe homozygote dans le gène *ESRP1* (NM_ 017697). Le score AR est de 650, pour un rang à 62. La protéine ESRP1 (*Epithelial splicing-regulatory protein 1*) est une protéine régulatrice de l'épissage spécifique des cellules épithéliales (Warzecha *et al*., 2009).

Une famille comprenant 2 enfants de sexes différents atteints de surdité de perception congénitale profonde a été étudiée récemment (Rohacek *et al*., 2017), retrouvant une hétérozygotie composite dans ce gène, avec un variant faux-sens hérité de la mère et une délétion de 19 pb héritée du père, entraînant un décalage du cadre de lecture et l'apparition prématuré d'un codon stop. Ce dernier codon intervient en amont de la séquence codante pour un domaine de reconnaissance du motif ARN, qui est essentiel à la fonction de la protéine. La présence d'apparentés sains, hétérozygotes pour l'une des deux mutations dans la famille, conduit à l'hypothèse d'une transmission autosomique récessive.

Dans cette même étude, un modèle d'embryons de souris *Esrp1*$^{-/-}$ a pu être étudié. La Figure 10 représente les résultats obtenus lors de cette étude fonctionnelle.

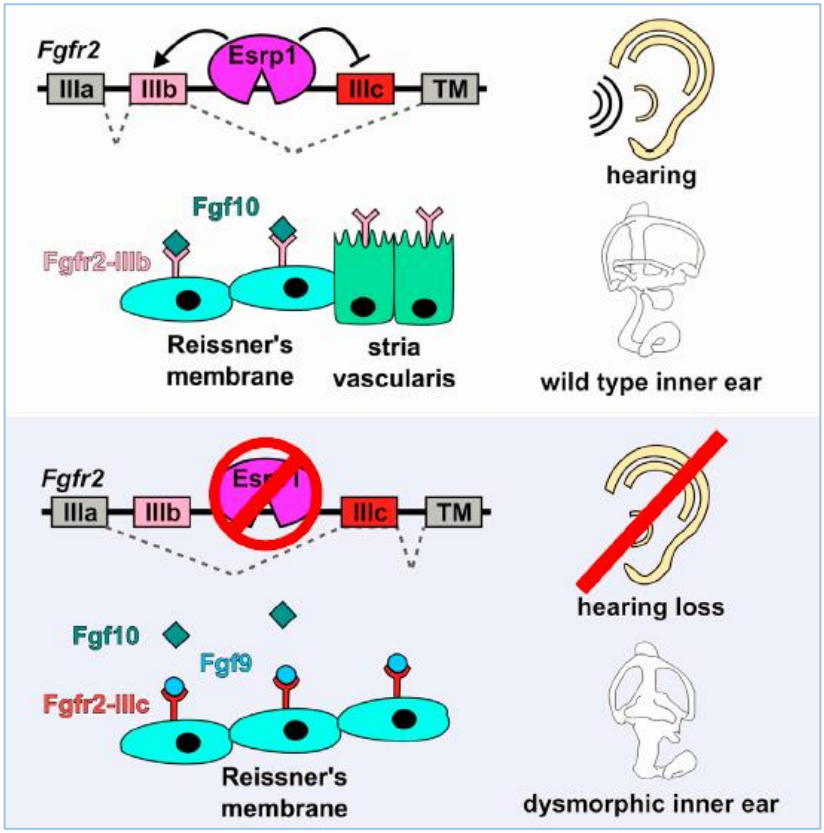


Figure 10. Etude fonctionnelle d'embryons de souris *Esrp1*$^{-/-}$ au niveau cochléaire (*d'après Rohacek* et al, *2017*)

En situation classique, l'expression du gène *ESRP1* est régulée de manière tissu-spécifique. Dans les cellules épithéliales cochléaires, la protéine Esrp1 participe à la régulation de l'épissage du gène *FGFR2*, induisant la production de l'isoforme protéique Fgfr2-IIIb, qui est l'isoforme épithélial. Son ligand spécifique est Fgf10 et leur liaison participe au développement des cellules de l'épithélium cochléaire, notamment celles de la membrane de Reissner. Dans les cellules mésenchymateuses où il n'y a pas d'expression d'*ESRP1*, l'épissage du gène *FGFR2* induit la production de l'isoforme protéique mésenchymateux Fgfr2-IIIc.

Dans ce modèle fonctionnel d'embryons de souris complétement délétées, le gène *ESRP1* n'est plus exprimé quel que soit le tissu. Il n'y a donc plus qu'une production de l'isoforme mésenchymateux Fgfr2-IIIc dans les cellules épithéliales. Le ligand spécifique de cet isoforme est Fgf9 et leur liaison participe aussi au développement et à l'expansion des cellules de la membrane de Reissner, mais au détriment du développement des cellules marginales, notamment des cellules de la stria vascularis. Les embryons étudiés présentent une anomalie de la morphogénèse de la cochlée, et donc de l'oreille interne, ce qui semble être à l'origine de la surdité.

Ceci pourrait expliquer la surdité profonde de ces patients, mais aussi celle que présente notre patient, issu de parents consanguins au second degré et sains, et dont la variation identifiée à l'état homozygote intervient également dans un domaine de reconnaissance du motif ARN.

Cet article précise aussi que l'étude de ces souris n'a pas pu être prolongée plus de quelques jours après leur naissance, puisqu'elles présentaient d'autres anomalies du développement embryonnaire telles des fentes labio-palatines et vélo-palatines qui ont conduit à leur décès. Ces fentes sont des embryopathies qui sont causées par un défaut de fusion des bourgeons de la face des bourgeons embryonnaires de la face pour les labio-palatines et des processus palatins pour les vélo-palatines.

Il est intéressant de considérer ces détails sachant que notre patient présente une luette bifide, qui est aussi une embryopathie intervenant sur le même axe de développement que les fentes vélo-palatines, c'est-à-dire qu'elle fait aussi suite aussi à un défaut de fusion des processus palatins.

Cependant, de nombreux facteurs peuvent être la cause de la luette bifide, tels des facteurs génétiques, environnementaux et toxiques. Il faut donc demeurer prudent quant à l'interprétation de cette caractéristique clinique chez notre patient et les fentes observées chez les souris *Esrp1*$^{-/-}$ étudiées.

L'étude de la transmission familiale présente un intérêt limité pour ce variant du fait de sa présence à l'état homozygote et de la consanguinité parentale : il est très probable que chacun de ses parents porte un exemplaire identique de cet allèle.

Par ailleurs, cette seule variation ne permet vraisemblablement pas d'expliquer la cardiopathie hypertrophique caractérisée chez ce patient. Il n'est pas non plus possible de déterminer si ce variant est directement impliqué dans la présentation d'un retard de développement.

5. Validité des scores établis pour l'identification de variants pathogènes

L'ensemble de règles du score AD est plutôt fiable bien qu'imparfait du point de vue diagnostique, au vu des scores supérieurs ou égaux à 350 obtenus dans l'étude rétrospective et de la détection de près de 92% des variants identifiés comme pathogènes dans les deux études combinées. La modulation de la pondération existante ou la pondération de nouvelles colonnes, notamment celles indiquant les annotations « Revel » et « Grantham », représentent des pistes intéressantes pour l'amélioration de ce score AD.

REVEL (*Rare Exome Variant Ensemble Learner*) est une méthode d'intégration de plusieurs outils de prédiction de pathogénicité des variations faux-sens (Ioannidis *et al.*, 2016). Parmi ces outils,

on retrouve notamment SIFT (*Sorting Intolerant From Tolerant*) ou PolyPhen (*Polymorphism Phenotyping*), qui évaluent l'impact potentiel provoqué par une substitution d'un acide aminé sur la stabilité et la fonction protéique, en comparant l'homologie de séquence et les propriétés physiques des acides. Le score Grantham, du nom de son créateur (Grantham, 1974), représente aussi une prédiction de l'effet des substitutions entre les acides aminés, cette fois-ci basée sur les propriétés physico-chimiques de ces acides, comme la polarité et le volume moléculaire.

Les variants pathogènes de transmission liée à l'X bénéficient d'un score X égal ou supérieur au seuil de 250 observé, soit une détection de 100% des variants pathogènes pour les 7 analyses d'exomes sur les deux études combinées. Ce score X demeure fiable, même s'il est nécessaire de tester sa validité sur un plus grand nombre d'exomes « positifs » avec un variant pathogène de transmission liée à l'X.

Malheureusement, le peu de variants pathogènes de transmission autosomique récessive ne permet pas de conclure sur la validité du score AR. Pour l'identification d'une pathologie issue d'une transmission hétérozygote composite, une des possibilités facilitant l'interprétation serait par exemple d'obtenir un tableur trié avec les variants d'un même gène regroupés en lignes successives, la ligne la plus haute de cet ensemble de lignes correspondant au meilleur score obtenu pour un des variants de ce gène. Ce n'est pas le cas actuellement, puisque le tableur est trié en fonction du score du variant seul, mais cela est envisagé par l'équipe du laboratoire dans un avenir proche.

6. Validité des scores pour la sélection de variants candidats dans la DI

Avec les ensembles de règles établies, la quasi-totalité des variants « recherche » mis en évidence se caractérise par une hypothèse de transmission autosomique dominante. Cela ne nous permet donc d'évaluer que la validité du score AD pour ces variants. Les scores obtenus pour les variations « recherche » d'intérêt sont compris entre 350 et 500.

Ces scores sont majoritairement proches de la valeur « seuil » pour le score AD, ce qui s'explique par des informations manquantes dans les bases de données, du fait que les gènes concernés ne soient pas répertoriés en pathologie humaine à l'heure actuelle. Mais il y a d'autres colonnes du tableur qui ne sont pas pondérées actuellement par le score AD et qui permettraient d'augmenter le score total attribué au variant, comme par exemple les colonnes mentionnées précédemment indiquant les annotations « Revel » et « Grantham ».

Il existe donc différentes manières d'améliorer le score AD établi pour mettre en évidence les variants « recherche » : soit pondérer positivement de nouvelles colonnes, contenant des informations sur la prédiction de l'impact du changement d'acide aminé, soit diminuer la pondération de la colonne « hypothèse de transmission » dans laquelle des informations sont absentes. Un autre moyen

d'améliorer ce score est de pondérer négativement le variant lorsque le mode de transmission est connu, car notre recherche porte sur des variants dont la transmission allélique est inconnue ou non encore reconnue pour la plupart d'entre eux.

7. Pertinence de l'analyse de l'exome dans le cadre de la recherche

L'analyse de l'exome est coûteuse (prix des réactifs) et nécessite une attention particulière pour le technicien et l'ingénieur qui s'occupent de l'analyse, puisque seules les étapes d'extraction et de séquençage proprement dites sont automatisées pour le moment au laboratoire. Le temps consacré à l'interprétation biologique est lui aussi non négligeable. De plus, elle nécessite une communication efficace entre les biologistes et les cliniciens.

Le délai d'attente du résultat est d'autant plus long pour les patients que le rendement diagnostic varie de 20 à 50% selon les séries (Lee *et al.*, 2014; Tan *et al.*, 2017*)* et donc qu'une grande partie des cas demeure non résolue à l'issue de l'analyse. En effet, l'analyse de l'exome est liée aux connaissances scientifiques, qui ne cessent d'évoluer dans le temps. La ré-analyse ponctuelle des données et des variants retrouvés augmente ce taux de rendement diagnostic, mais aucune recommandation n'est actuellement faite sur l'intervalle de temps entre l'analyse initiale des données et leur ré-analyse.

Pour des raisons économiques, le laboratoire a choisi de réaliser des analyses d'exome en simplex et non en trio. L'étude de la transmission familiale des variants retenus est ensuite réalisée par séquençage Sanger. L'analyse d'exomes en trio, hormis la question d'un coût plus onéreux, a l'avantage d'inclure les données de la ségrégation familiale d'emblée et donc de pouvoir mettre en avant les variations survenues *de novo*. Pour le moment, le laboratoire n'opte pour cette stratégie d'analyse en trio qu'en cas d'analyse simplex non concluante.

Une alternative à l'analyse d'exomes en trio est la constitution de « pools » parentaux, c'est-à-dire d'échantillons parentaux mélangés de façon équimolaire afin de construire une seule librairie. Ces mélanges permettent de retrouver une fraction de chaque variant parental dans l'analyse et de comparer ces variants à ceux présentés par le cas index. On peut ainsi identifier les variations héritées et les variations *de novo*.

Pour pouvoir analyser correctement les données des variants issues de ces fractions d'ADN parentaux, nous sélectionnons pour chaque mélange 4 parents au maximum du même sexe. En effet, au-delà de ce nombre, la fraction parentale se révèle parfois trop faible pour obtenir une couverture de séquençage et une profondeur de lecture suffisantes des variants présents chez les parents du cas

index. L'identification de ces variants parentaux n'est alors pas possible ce qui entraîne la détection de variants faux positifs *de novo* chez leur enfant.

Toutefois, si cette solution est économique par rapport à l'analyse de l'exome en trio du point de vue de l'utilisation des réactifs et de la *flow cell*, la problématique principale reste la disponibilité des prélèvements parentaux. Cette situation tend à se raréfier, puisque l'étude de la transmission allélique dans les analyses menées au laboratoire est systématique et donc que les cliniciens demandent d'emblée les prélèvements des deux parents du cas index lors de la consultation.

Une autre problématique de cette méthode concerne l'extraction de la fraction de ces variants parentaux. Pour le moment, celle-ci n'est pas automatisée et nécessite une analyse des données informatiques rigoureuse variant par variant, ce qui demande un temps précieux.

L'analyse de ces mélanges parentaux est disponible pour 7 des patients inclus dans l'étude prospective. Elle a permis de mettre en évidence des variants *de novo* pour 4 d'entre eux, qui malheureusement n'ont pas pu être étudiés dans ce projet par manque de temps. Ces variants ainsi extraits constituent une piste d'exploration intéressante à l'avenir pour expliquer le phénotype de ces patients.

8. Limites de la technique de l'exome dans l'analyse pangénomique

Telle que construite pour notre projet, l'analyse des données de l'exome se limite à l'interprétation des variations nucléotidiques. Les logiciels utilisés dans cette étude pour l'analyse bio-informatique ne comprennent pas d'outil informatique spécifique au traitement des données et à l'interprétation des CNV. Il est donc nécessaire de réaliser une analyse pangénomique en amont de l'analyse de l'exome, telle l'ACPA, pour identifier ces CNV.

En effet, le caryotype moléculaire reste une technique diagnostique très résolutive, plus rapide et bien moins coûteuse que l'exome, qui permet de mettre en évidence, par exemple, des délétions de très petites tailles pouvant expliquer à elles seules le phénotype clinique du patient. A l'heure actuelle, l'ACPA et l'exome restent donc deux techniques complémentaires en matière de diagnostic. Le développement d'un logiciel permettant d'identifier les CNV dont la taille n'est pas détectable par l'ACPA est en cours.

Outre son coût économique trop important actuellement pour les laboratoires de diagnostic, l'analyse du génome entier présente un réel intérêt à l'avenir car elle permet la détection non seulement de variations nucléotidiques dans le génome entier, mais aussi de réarrangements

chromosomiques de petites tailles, même ceux qui sont équilibrés (translocation et inversion chromosomiques).

De plus, il n'est pas utile d'effectuer des étapes de capture de l'ADN pour cette technique, puisque c'est le génome dans sa globalité qui est analysé. Une étude récente a montré que ces particularités techniques permettent une meilleure couverture des régions codantes, notamment celles riches en GC, avec une plateforme de séquençage du génome entier qu'avec une plateforme de séquençage de l'exome (Meienberg *et al.*, 2016).

De plus, les maladies génétiques à expansion de tri-nucléotides peuvent être également décelées (Tang *et al*., 2017). Il est donc possible que dans un avenir plus ou moins proche l'analyse du génome entier soit réalisée en première intention.

V. Conclusion

L'analyse de l'exome est aujourd'hui primordiale afin d'identifier des variations géniques pathogènes avec un niveau de résolution à l'échelle du nucléotide et d'avancer dans le diagnostic de certaines pathologies, comme la DI. A l'avènement de l'ère pangénomique, il est nécessaire de développer des outils informatiques, tels l'application GenSCor, pour améliorer l'efficacité dans l'interprétation biologique du très grand nombre de variants générés par l'analyse. Cette application permet d'attribuer un score à chaque variant, dont l'ensemble des règles établies est adapté à l'hypothèse de transmission allélique.

Au vu des nombreuses annotations disponibles pour chacun des variants, il est indispensable d'adapter la pondération de chaque règle pour obtenir un score fiable permettant une interprétation juste et relativement rapide des données recueillies informatiquement. Il faut cependant garder à l'esprit que l'annotation d'un variant dépend de nombreux outils informatiques, qui, à l'heure actuelle, ne peuvent malheureusement pas progresser aussi rapidement que l'évolution des connaissances. Ceci peut générer des erreurs pendant l'étape d'annotation des variants, qui se répercutent lors de l'attribution d'un score.

Dans ce projet, trois scores AD, AR et X ont été établis, mais seule la validité du score AD a pu être attestée par l'identification de 92% des variations « pathogènes » dans les 2 études. L'analyse des données du score AD a permis de mettre en évidence une valeur « seuil » en dessous de laquelle il n'est plus nécessaire de considérer les variants pour l'interprétation de l'analyse. Par ailleurs, le rendement diagnostic dans l'étude prospective est de 16%, ce qui devrait évoluer ces prochaines années grâce aux progrès techniques et l'évolution constante des connaissances en génétique.

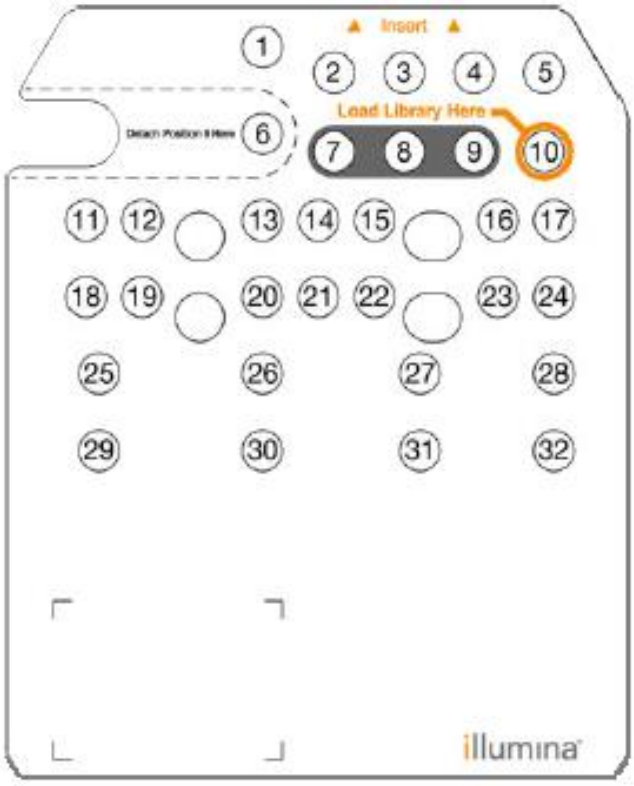
Dans le cadre de la recherche, nous avons sélectionné des variants candidats avec une hypothèse de transmission AD. L'annotation de ces variants et l'étude de la ségrégation familiale, même

complète, ne permettent pas encore de conclure sur la pathogénicité potentielle de ces variants. Cette analyse a néanmoins permis de mettre en évidence des gènes candidats pour la DI, *USP19*, *FKBP4* et *NCKAP1*, pour lesquels des études fonctionnelles sont actuellement en cours, grâce à des collaborations internationales. De ce point de vue, nous pouvons dire que l'analyse de l'exome se situe à la frontière entre le cadre diagnostic et la recherche.

VI. Annexes

Annexe 1: Cartouche de réactifs (*d'après une image Illumina*)

préremplie de réactifs d'amplification et de séquençage.



- Puits 7, 8 et 9 pour le chargement respectivement des solutions de réactifs 20, 21 et 22 avec ajout des amorces spécifiques (séquençage « *paired-end* »)
- Puits 10 pour le chargement du mélange des librairies et du marqueur PhiX

Annexe 2 : Extrait de tableur Excel issu du fichier .VCF, avec les annotations principales

Score	ID_MT	Gene	cNomen	pNomen	Haplotype	Effet_mutation	Nbre_allele	Allèle_GnomAD	Couverture	Presence_pc	Phred_score
700	chr15:75693133G>A	SIN3A	c.1675C>T	p.Arg559*	het	stop_gained	1	0	86	44	904.77
550	chr1:156096670C>A	LMNA	c.77C>A	p.Ala26Asp	het	missense_variant	1	0	105	53	1256.77
550	chr5:67589301A>G	PIK3R1	c.1289A>G	p.Lys430Arg	het	missense_variant	1	3	35	63	603.77
450	chr12:49434627G>A	KMT2D	c.6926C>T	p.Ser2309Leu	het	missense_variant	1	2	380	42	3622.77
400	chr12:8995728A>C	A2ML1	c.1249-2A>C	.	het	splice_acceptor_variant	1	0	126	44	1357.77
400	chr14:63858549delAT	PPP2R5E	c.939_940delAT	p.Cys314fs	het	frameshift_variant	1	0	15	60	322.73
400	chr11:68696725C>T	IGHMBP2	c.1135C>T	p.Gln379*	het	stop_gained	1	0	224	54	2966.77

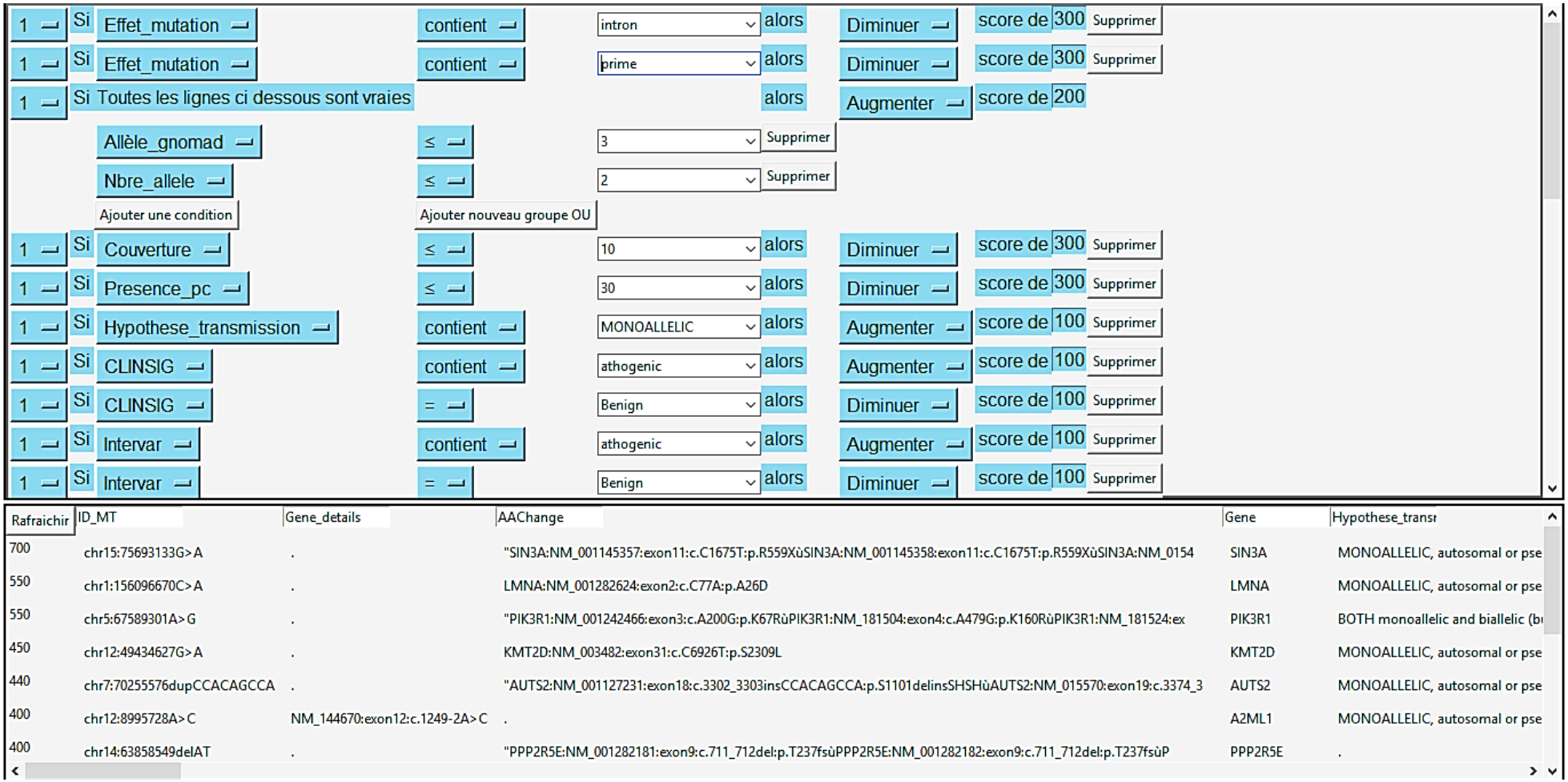
Hypothese_transmission	Domino_score	CLINSIG	Intervar	role_DI	Revel	Grantham	dbscSNV_ADA	DANN_pred	Sift_pred	Provean_pred	Polyphen_HDIV
MONOALLELIC	0.999700385912	Pathogenic	Likely pathogenic	oui(Vert)	.	.	.	0.880	.	.	.
MONOALLELIC	0.998895177259	.	Uncertain significance	oui(Orange)	0.171	126	.	0.240	T	N	B
BOTH monoallelic and biallelic	0.999235981587	.	Uncertain significance	oui(Orange)	0.236	26	.	0.272	T	N	B
MONOALLELIC	0.997706636720	.	Uncertain significance	oui(Vert)	0.318	145	.	0.271	D	N	B
MONOALLELIC	0.092256111995	.	.	oui(Rouge)	.	.	1.0000	0.229	.	.	.
.	0.860174227747	.	.	.	.	.	.	.	.	.	.
BIALLELIC	0.071846022235	.	Likely pathogenic	oui(Orange)	.	.	.	0.912	.	.	.

LRT_pred	Mutation_taster	FATHMM_pred	Meta_SVP_pred	Meta_LR
D	D	.	.	.
.	N	D	T	T
D	D	T	T	T
N	N	T	T	T
.	D	.	.	.
.	.	.	.	.
D	D	.	.	.

D : délétère ; T : toléré ; N : neutre ; B : bénin ; le point symbolise une information manquante dans les bases de données, a priori non connue

La colonne Score est indiquée en violet. L'identité du variant et sa localisation sont disponibles dans les colonnes jaunes. La fréquence allélique est évaluée dans les colonnes en vert, dans les exomes analysés au laboratoire jusqu'à présent (colonne « Nbre_allele ») et dans la base de données GnomAD (« Allèle_GnomAD »). Les colonnes en bleu contiennent des informations sur la qualité du séquençage, les colonnes en rouge sur l'hypothèse de transmission et celles en orange la classification si ce variant est connu dans des bases de données. Les colonnes en gris représentent les données de prédiction établies par différents scores d'intégration et outils informatiques de prédiction.

Annexe 3 : Application GenSCor, avec un extrait de l'ensemble de règles du score AD



Le premier bloc correspond à l'ensemble des règles qui permettent d'attribuer un score aux variants. Le deuxième correspond au résultat obtenu grâce à l'application ; le score apparaît en 1ere position, suivi par l'identité du variant et son annotation complète.

VII. Références bibliographiques

Anazi, S., Maddirevula, S., Salpietro, V., Asi, Y.T., Alsahli, S., Alhashem, A., Shamseldin, H.E., AlZahrani, F., Patel, N., Ibrahim, N., et al. (2017). Expanding the genetic heterogeneity of intellectual disability. Hum. Genet. *136*, 1419–1429.

Bai, X., Edwards, J., and Ju, J. (2005). Molecular engineering approaches for DNA sequencing and analysis. Expert Rev. Mol. Diagn. *5*, 797–808.

Berry-Kravis, E., Abrams, L., Coffey, S.M., Hall, D.A., Greco, C., Gane, L.W., Grigsby, J., Bourgeois, J.A., Finucane, B., Jacquemont, S., et al. (2007). Fragile X-associated tremor/ataxia syndrome: clinical features, genetics, and testing guidelines. Mov. Disord. Off. J. Mov. Disord. Soc. *22*, 2018–2030, quiz 2140.

Bloch, J., Cans, C., de Vigan, C., de Brosses, L., Doray, B., Larroque, B., and Perthus, I. (2008). Faisabilité de la surveillance du syndrome d'alcoolisation fœtale (SAF). Arch. Pédiatrie *15*, 507–509.

Chamberlain, S.J., and Lalande, M. (2010). Angelman syndrome, a genomic imprinting disorder of the brain. J. Neurosci. Off. J. Soc. Neurosci. *30*, 9958–9963.

Cogné, B., Ehresmann, S., Beauregard-Lacroix, E., Rousseau, J., Besnard, T., Garcia, T., Petrovski, S., Avni, S., McWalter, K., Blackburn, P.R., et al. (2019). Missense Variants in the Histone Acetyltransferase Complex Component Gene TRRAP Cause Autism and Syndromic Intellectual Disability. Am. J. Hum. Genet. *104*, 530–541.

Colantuoni, C., Purcell, A.E., Bouton, C.M., and Pevsner, J. (2000). High throughput analysis of gene expression in the human brain. J. Neurosci. Res. *59*, 1–10.

Croen, L.A., Grether, J.K., and Selvin, S. (2001). The epidemiology of mental retardation of unknown cause. Pediatrics *107*, E86.

Dave, U., Shetty, N., and Mehta, L. (2005). A community genetics approach to population screening in India for mental retardation--a model for developing countries. Ann. Hum. Biol. *32*, 195–203.

David, M., Dieterich, K., Billette de Villemeur, A., Jouk, P.-S., Counillon, J., Larroque, B., Bloch, J., and Cans, C. (2014). Prevalence and characteristics of children with mild intellectual disability in a French county. J. Intellect. Disabil. Res. JIDR *58*, 591–602.

Eldomery, M.K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J.A., Gambin, T., Stray-Pedersen, A., Küry, S., Mercier, S., Lessel, D., Denecke, J., et al. (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. *9*, 26.

El-Saafin, F., Curry, C., Ye, T., Garnier, J.-M., Kolb-Cheynel, I., Stierle, M., Downer, N.L., Dixon, M.P., Negroni, L., Berger, I., et al. (2018). Homozygous TAF8 mutation in a patient with intellectual disability results in undetectable TAF8 protein, but preserved RNA polymerase II transcription. Hum. Mol. Genet. *27*, 2171–2186.

Emerson, E. (2012). The World Report on Disability. J. Appl. Res. Intellect. Disabil. *25*, 495–496.

Fannemel, M., Barøy, T., Holmgren, A., Rødningen, O.K., Haugsand, T.M., Hansen, B., Frengen, E., and Misceo, D. (2014). Haploinsufficiency of XPO1 and USP34 by a de novo 230 kb deletion in 2p15, in a patient with mild intellectual disability and cranio-facial dysmorphisms. Eur. J. Med. Genet. *57*, 513–519.

Fattorini, G., Antonucci, F., Menna, E., Matteoli, M., and Conti, F. (2015). Co-expression of VGLUT1 and VGAT sustains glutamate and GABA co-release and is regulated by activity in cortical neurons. J. Cell Sci. *128*, 1669–1673.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. Science *185*, 862–864.

Gurrieri, F., and Accadia, M. (2009). Genetic imprinting: the paradigm of Prader-Willi and Angelman syndromes. Endocr. Dev. *14*, 20–28.

Gustavson, K.-H. (2005). Prevalence and aetiology of congenital birth defects, infant mortality and mental retardation in Lahore, Pakistan: A prospective cohort study. Acta Paediatr. *94*, 769–774.

Hassink, G.C., Zhao, B., Sompallae, R., Altun, M., Gastaldello, S., Zinin, N.V., Masucci, M.G., and Lindsten, K. (2009). The ER-resident ubiquitin-specific protease 19 participates in the UPR and rescues ERAD substrates. EMBO Rep. *10*, 755–761.

Hausch, F. (2015). FKBPs and their role in neuronal signaling. Biochim. Biophys. Acta *1850*, 2035–2040.

Hoyme, H.E., May, P.A., Kalberg, W.O., Kodituwakku, P., Gossage, J.P., Trujillo, P.M., Buckley, D.G., Miller, J.H., Aragon, A.S., Khaole, N., et al. (2005). A Practical Clinical Approach to Diagnosis of Fetal Alcohol Spectrum Disorders: Clarification of the 1996 Institute of Medicine Criteria. Pediatrics *115*, 39–47.

International Human Genome Sequencing Consortium, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860.

Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

Larroque, B., Ancel, P.-Y., Marret, S., Marchand, L., André, M., Arnaud, C., Pierrat, V., Rozé, J.-C., Messer, J., Thiriez, G., et al. (2008). Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the EPIPAGE study): a longitudinal cohort study. Lancet Lond. Engl. *371*, 813–820.

Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA *312*, 1880–1887.

Leonard, H., and Wen, X. (2002). The epidemiology of mental retardation: Challenges and opportunities in the new millennium. Ment. Retard. Dev. Disabil. Res. Rev. *8*, 117–134.

Lévy, J., Coussement, A., Dupont, C., Guimiot, F., Baumann, C., Viot, G., Passemard, S., Capri, Y., Drunat, S., Verloes, A., et al. (2017). Molecular and clinical delineation of 2p15p16.1 microdeletion syndrome. Am. J. Med. Genet. A. *173*, 2081–2087.

Lui, T.T.H., Lacroix, C., Ahmed, S.M., Goldenberg, S.J., Leach, C.A., Daulat, A.M., and Angers, S. (2011). The ubiquitin-specific protease USP34 regulates axin stability and Wnt/β-catenin signaling. Mol. Cell. Biol. *31*, 2053–2065.

Lupski, J.R. (2010). New mutations and intellectual function. Nat. Genet. *42*, 1036–1038.

Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: A meta-analysis of population-based studies. Res. Dev. Disabil. *32*, 419–436.

May, P.A., Fiorentino, D., Coriale, G., Kalberg, W.O., Hoyme, H.E., Aragón, A.S., Buckley, D., Stellavato, C., Gossage, J.P., Robinson, L.K., et al. (2011). Prevalence of children with severe fetal alcohol spectrum disorders in communities near Rome, Italy: new estimated rates are higher than previous estimates. Int. J. Environ. Res. Public. Health *8*, 2331–2351.

Meienberg, J., Bruggmann, R., Oexle, K., and Matyas, G. (2016). Clinical sequencing: is WGS the better WES? Hum. Genet. *135*, 359–362.

Redin, C., Gérard, B., Lauer, J., Herenger, Y., Muller, J., Quartier, A., Masurel-Paulet, A., Willems, M., Lesca, G., El-Chehadeh, S., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. J. Med. Genet. *51*, 724–736.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–423.

Roeleveld, N., and Zielhuis, G.A. (1997). The prevalence of mental retardation: a critical review of recent literature. Dev. Med. Child Neurol. *39*, 125–132.

Rohacek, A.M., Bebee, T.W., Tilton, R.K., Radens, C.M., McDermott-Roe, C., Peart, N., Kaur, M., Zaykaner, M., Cieply, B., Musunuru, K., et al. (2017). ESRP1 Mutations Cause Hearing Loss due to Defects in Alternative Splicing that Disrupt Cochlear Development. Dev. Cell *43*, 318-331.e5.

Rousseau, T., Amar, E., Ferdynus, C., Thauvin-Robinet, C., Gouyon, J.-B., and Sagot, P. (2010). Variations de prévalence de la trisomie 21 en population française entre 1978 et 2005. J. Gynécologie Obstétrique Biol. Reprod. *39*, 290–296.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. Hum. Mutat. *36*, 928–930.

Srour, M., and Shevell, M. (2014). Genetics and the investigation of developmental delay/intellectual disability. Arch. Dis. Child. *99*, 386–389.

Steffen, A., Rottner, K., Ehinger, J., Innocenti, M., Scita, G., Wehland, J., and Stradal, T.E. (2004). Sra-1 and Nap1 link Rac to actin assembly driving lamellipodia formation. EMBO J. *23*, 749–759.

Tan, T.Y., Dillon, O.J., Stark, Z., Schofield, D., Alam, K., Shrestha, R., Chong, B., Phelan, D., Brett, G.R., Creed, E., et al. (2017). Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions. JAMA Pediatr. *171*, 855–862.

Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., et al. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am. J. Hum. Genet. *101*, 700–715.

Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2016). Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. *17*, 9–18.

Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B., and Carstens, R.P. (2009). ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. Mol. Cell *33*, 591–601.

Zhang, X., Ling, J., Barcia, G., Jing, L., Wu, J., Barry, B.J., Mochida, G.H., Hill, R.S., Weimer, J.M., Stein, Q., et al. (2014). Mutations in QARS, Encoding Glutaminyl-tRNA Synthetase, Cause Progressive Microcephaly, Cerebral-Cerebellar Atrophy, and Intractable Seizures. Am. J. Hum. Genet. *94*, 547–558.

UNIVERSITE DE POITIERS

Faculté de Médecine et de Pharmacie

SERMENT

⌘⌖⌘⌖⌘

En présence des Maîtres de cette école, de mes chers condisciples et devant l'effigie d'Hippocrate, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité dans l'exercice de la médecine. Je donnerai mes soins gratuits à l'indigent et n'exigerai jamais un salaire au-dessus de mon travail. Admise dans l'intérieur des maisons mes yeux ne verront pas ce qui s'y passe ; ma langue taira les secrets qui me seront confiés, et mon état ne servira pas à corrompre les mœurs ni à favoriser le crime. Respectueuse et reconnaissante envers mes Maîtres, je rendrai à leurs enfants l'instruction que j'ai reçue de leurs pères.

Que les hommes m'accordent leur estime si je suis fidèle à mes promesses ! Que je sois couverte d'opprobre et méprisée de mes confrères si j'y manque !

⌘⌖⌘⌖⌘

RESUMÉ

De nos jours, la déficience intellectuelle (DI) représente une question importante de santé publique. La prévalence de la DI est estimée à 1% de la population mondiale et la plupart des causes reste encore inconnue actuellement. Parmi elles, les causes génétiques représentent 15 à 50%. Grâce aux innovations technologiques comme le développement des techniques de séquençage haut débit, il nous est aujourd'hui possible d'identifier de nouveaux gènes impliqués dans la DI, par l'analyse de l'exome, la partie codante du génome.

Pour ce projet, nous avons réalisé deux types d'études en parallèle. Dans une étude rétrospective, nous avons analysé 21 exomes « contrôles » dans lesquels une mutation pathologique a déjà été identifiée. Cette première analyse nous a permis d'élaborer trois métascores, issus de trois ensembles de règles. Chaque ensemble correspond à une pondération des règles différente selon l'hypothèse de transmission allélique prise en compte pour établir le score. L'élaboration de ces scores doit permettre une meilleure efficacité dans le tri des variants issus des données de l'analyse.

Dans une étude prospective, nous avons réalisé l'analyse de l'exome de 62 patients présentant une DI ou un retard des apprentissages à l'aide des scores établis, avec un rendement diagnostic de 16%. Nous avons identifié de plus dans trois gènes, *USP19*, *NCKAP1* et *FKBP4*, des variants « recherche » candidats dans la DI, pour lesquels des études fonctionnelles sont en cours à l'heure actuelle grâce à des collaborations internationales.

Mots clés : séquençage haut débit ; exome ; gène ; déficience intellectuelle ; retard développemental ; métascore ; FKBP4 ; NCKAP1 ; USP19