Approches bioinformatiques innovantes pour l'analyse de données de séquençage à haut-débit appliquées à l'étude de pathologies génétiques rares avec anomalies du développement

Philippine Garret

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
PRÉPARÉE À l'unité INSERM 1231 – Équipe GAD

École doctorale n° 554
« Environnements-Santé »

Doctorat de Biochimie et biologie moléculaire

Par
Philippine GARRET

Approches bioinformatiques innovantes pour l'analyse de données de séquençage à haut débit appliquées à l'étude de pathologies génétiques rares avec anomalies du développement

Thèse présentée et soutenue à Dijon, le 27 mai 2020

Composition du Jury :

Pr Christophe PHILIPPE	PU-PH, Université de Bourgogne Franche Comté	Président du jury
Dr Nicolas THIERRY-MIEG	Chargé de Recherche, Université Grenoble Alpes	Rapporteur
Dr Gaël NICOLAS	MCU-PH, CHU Rouen	Rapporteur
Dr Gaetan LESCA	MCU-PH, Groupement Hospitalier Est Bron	Examinateur
Pr Laurence FAIVRE	PU-PH, Université de Bourgogne Franche comté	Directrice de thèse
Pr Christel THAUVIN	PU-PH, Université de Bourgogne Franche comté	Codirectrice de thèse
Dr Deltef TROST	Cytogénéticien, Laboratoire Cerba	Invité
Mr Yannis DUFFOURD	Ingénieur de Recherche, CHU Dijon	Invité

A ma famille

REMERCIEMENTS

Cette thèse a été réalisée à l'Université de Bourgogne Franche-Comté dans le cadre d'une convention CIFRE entre le Laboratoire de Génétique des Anomalies du Développement (GAD) à Dijon, la Région Bourgogne Franche-Comté et le Laboratoire Cerba à Saint-Ouen l'Aumône.

Je tiens tout d'abord à remercier Laurence Faivre et Christel Thauvin-Robinet de m'avoir accueillie au sein de l'équipe GAD et d'avoir accepté de diriger mes travaux. Merci d'avoir permis à cette thèse CIFRE d'avoir lieu dans les meilleures conditions. Merci pour vos conseils avisés, votre disponibilité et vos encouragements.

Je remercie aussi Jean-Marc Costa, Detlef Trost et Aïcha Boughalem de m'avoir accueillie au Laboratoire Cerba. C'était toujours un plaisir de travailler et échanger avec vous.

J'exprime également ma gratitude envers différents membres de mon jury de thèse pour avoir accepté d'évaluer mes travaux. Je remercie Christophe Philippe d'avoir accepté d'être président du jury. Je suis reconnaissante envers Gaetan Lesca pour son implication dans l'évaluation de mes travaux. Je remercie Nicolas Thierry-Mieg et Gaël Nicolas d'avoir consacré du temps à la lecture de ce manuscrit.

Je souhaite remercier Yannis Duffourd de m'avoir aussi bien encadrée durant ces trois ans sur la partie bioinformatique et plus largement la thèse. Merci à Emilie Tisserant, Anne-Sophie Denommé-Pichon et Simon Verdez pour vos précieux conseils en la matière.

Je suis reconnaissante envers Jean Muller et Julien Thevenon d'avoir fait partie de mon comité de suivi de thèse.

Je tiens à remercier les collaborateurs extérieurs qui ont participé activement à ces différents projets et grâce à qui mes travaux ont pu être pu être publiés. Je remercie également les patients et leur famille.

Je remercie Ange-Line Bruel d'avoir pris sous son aile la petite nouvelle et de m'avoir montré les ficelles du doctorant. C'est précieux lorsqu'on arrive en cours d'année ! Merci pour la formation exome et pour ta patience face à mes (très!) nombreuses questions. Merci également

d’avoir organiser ces sorties qui ont soudé l’équipe.

Je souhaite également remercier Antonio Vitobello pour le temps et le soutien qu’il m’a accordés pour mener à bien mes travaux, plus particulièrement la publication sur *OTUD7A*. Merci aussi à Laurence Jego pour son implication dans les manips de fonctionnel que je ne pouvais pas réaliser.

Merci aux membres de l’équipe GAD :
Aux chercheurs, ingénieurs et post-docs aux spécialités de plus en plus diversifiées. Merci pour ces conseils et les discussions très variées que nous avons pu avoir.
Aux techniciens Martin, Thibaud, Charlotte, Morgane, Valentin, Victor et Sylvie pour leur aide précieuse pendant les manips et pour m’avoir permis de renouer avec la paillasse. À Martin pour son implication essentielle dans le projet « éléments mobiles ».
À mes co-thésards, aux étudiants, aux internes et aux stagiaires pour votre bonne humeur.

Un grand merci à Lionel, Coralie et les membres de l’Expérimentarium. Ces coupures dans la vie trépidante de thésarde redonnent toujours de la motivation pour avancer.

Un immense merci à mes proches :
À mes amis, les éternels Illustratrice (vive les grenouilles), Scientifiques, Retardataires et autres Surdoués pour ces bons moments passés en cours (!) et en dehors.
À Mathilde et Bérénice et leur +1 pour leur soutien durant toutes ces années, de m’avoir permis d’être cheftaine et de m’avoir fait visiter la France et les mairies de Paris ! Merci d’avoir pris régulièrement des nouvelles de la provinciale.
À mon oncle et mes tantes pour leur accueil parisien, dijonnais ou beaunois et leur écoute depuis toutes ces années.
À mes grands-parents, ma famille et aux nouveaux membres qui vont nous rejoindre d’ici ma soutenance.
À mon frère et ma sœur pour leurs encouragements et leur brin d’humour qui donnent le sourire dès le matin.
À mes parents pour avoir toujours été là pour moi, pour m’avoir écoutée parler de biologie pendant toutes ces années (ce n’est pas fini !) et pour m’avoir supportée ces 6 derniers mois, confinement inclus !

Enfin un grand merci à ceux tous ceux qui ont été présents durant ces 3 ans.

RÉSUMÉ

L'avènement du séquençage haut débit d'exome (ES) en diagnostic et en recherche ces dernières années a conduit à l'identification des bases génétiques de nombreuses pathologies mendéliennes, permettant de résoudre de nombreuses situations d'errance diagnostique. Néanmoins, l'analyse des données de ES permet uniquement d'identifier des variations pathogènes ou probablement pathogènes dans 30 à 45 % des situations sans diagnostic. En effet, certaines limites existent, tant au niveau clinique, moléculaire et bioinformatique. L'évolution constante des connaissances cliniques, du nombre de nouveaux gènes impliqués en pathologie humaine, et des corrélations clinico-biologiques a un impact important sur l'analyse des données, entraînant une amélioration progressive de la recherche diagnostique. Des limites techniques inhérentes à la technologie, avec en particulier des régions non couvertes, existent, mais se sont également significativement réduites ces dernières années. Enfin, au-delà de l'analyse de SNV et de CNV, d'autres anomalies génétiques peuvent être responsables de maladies rares, nécessitant un développement bioinformatique pour optimiser les résultats. Bien que le séquençage à haut débit du génome permette de résoudre des observations, en particulier en cas de variations dans les régions non codantes ou les variations de structure, il existe encore de nombreuses informations à extraire et à exploiter à partir des données de ES.

L'objectif de cette thèse a donc été de participer à l'amélioration des approches bioinformatiques d'analyse de données de ES pour l'identification de nouveaux gènes ou mécanismes moléculaires impliqués dans des maladies génétiques rares afin de réduire l'errance diagnostique des patients.

Plusieurs stratégies ont ainsi été mises en place. La première stratégie a consisté en une réanalyse recherche de données de 80 patients ayant bénéficié d'un ES au laboratoire CERBA (thèse CIFRE) dont la lecture diagnostique était négative. Elle a conduit à la mise en évidence deux nouveaux gènes candidats dans la déficience intellectuelle syndromique, dont le gène *OTUD7A* (article 1). La deuxième stratégie a consisté en la mise au point d'un pipeline bioinformatique pour extraire les données du génome mitochondrial à partir des données de ES. L'ADN mitochondrial n'est pas ciblé par les kits de capture d'exome mais peut être extrait des données capturées indirectement, rendant son analyse possible à partir de données de ES préexistantes. A partir de la collection GAD d'exomes de patients sans diagnostic, deux variations causales ont été identifiées chez deux individus atteints de troubles neuro-développementaux sur 928 personnes étudiées, et ainsi résoudre une errance diagnostique dans 0,2 % des patients sans diagnostic (article 2). La troisième stratégie a consisté en la mise en place d'un pipeline bioinformatique d'identification des éléments mobiles au sein des données d'exome, étant attendu qu'environ 0,3 % des variations pathogènes du génome humain ont pour origine l'insertion *de novo* d'un élément mobile. A partir de la collection GAD d'exomes de 3322 individus (2500 cas index) sans diagnostic, cette étape a permis d'identifier deux cas solides en lien avec l'insertion d'un élément Alu au sein d'un exon du gène *FERMT1* et du gène *GRIN2B* (article 3 en cours d'écriture).

Cette thèse a permis de repousser certaines limites de la technologie d'exome. D'autres perspectives existent, et sont explorées par l'équipe, en lien avec le projet Européen Solve-RD.

ABSTRACT

In recent years, the advent of exome sequencing (ES) in diagnosis and research has led to the identification of the genetic bases of many Mendelian disorders. In turn, this has allowed researchers to resolve many undiagnosed cases. Nevertheless, ES data analysis only leads to the identification of pathogenic or likely pathogenic variants in 30 to 45 % of these unsolved cases, which are known as diagnostic odysseys. Indeed, There are some limitations at the clinical, molecular and bioinformatics levels. The constant progression of clinical knowledge, of the number of genes involved in human diseases, and of clinical-biological correlations have had a significant impact on data analysis, leading to a progressive improvement in diagnostic research. However there are limits to the current technologies, particularly for regions that are not covered, though these limits have been significantly reduced over the last few years. Although genome sequencing will certainly resolve a number of undiagnosed cases, especially in case of non-coding or structural variants, there is still a lot information to be extracted and analyzed from ES data. Finally, beyond SNV and CNV analyses, other genetic events can be involved in rare disorders, requiring the development of bioinformatics to optimize results.

The aim of this project was therefore to improve bioinformatics approaches to ES data analysis in order to identify new molecular mechanisms involved in rare genetic disorders and thus reduce the number of diagnostic odysseys.

Several strategies were established. The first consisted in reanalyzing ES data from 80 undiagnosed patients who were sequenced by the Laboratoire CERBA (CIFRE thesis). This led to the identification of 2 new candidate genes involved in ID, particularly the *OTUD7A* gene (Article 1). The second strategy was to develop a bioinformatics pipeline in order to extract mitochondrial DNA data from ES data. The mitochondrial genome is not targeted by exome capture kits but can be extracted from off-target data, which provides an opportunity to analyze it from preexisting ES data. From the GAD exomes cohort of undiagnosed patients, 2 causal variations were identified in 2 out of 928 individuals with a neuro-developmental disorder. This approach therefore resolved the diagnostic odyssey of 0.2 % of undiagnosed patients (Article 2). The third strategy was to develop a bioinformatics pipeline to identify the insertion of mobile elements within ES data, with the expectation that about 0.03 % of pathogenic variants originate from *de novo* insertion of mobile elements. This step lead to the identification of 2 cases of Alu element insertion in *FERMT1* and *GRINB2* gene exons from the GAD exomes cohort of 3322 undiagnosed individuals (2500 probands) (Article 3, in progess).

The research undertaken for this PhD has expanded some of the limits of ES. Still other perspectives exist and are currently being explored by the GAD team in collaboration with the European Solve-RD project.

ABRÉVIATIONS

AD : Anomalie du Développement
ADAR : Adenosine Deaminase Acting on RNA
ADNc : ADN complémentaire
ADNmt : ADN mitochondrial
ARNm : ARN messager
ASP : AntiSens Promoter
ATP : Adenosine Triphosphate
BAM : Binary Alignment/Map
BWA : Burrows-Wheeler Aligner
CADD : Combined Annotation Dependent Depletion
cassure DB : cassure double-brin
CHG : Comparative Genomic Hybridization
CNV : Copy Number Variation
COF : Cortex OrbitoFrontal
dbSNP : Single Nucleotide Polymorphism Database
DDD : Deciphering Developmental Disorders
dNTP : désoxyribonucléoside triphosphate
DGV : Database of Genomic Variants
DI : Déficience Intellectuelle
DP : Paire de lectures discordantes
DUB : DéUBiquitinase
EEG : ElectroEncéphaloGraphie
EM : Élément mobile
ES : Séquençage d'exome
ET : Élément Transposable
GATK : Genome Analysis ToolKit
GERP : Genomic Evolutionary Rate Profiling
GS : Séquençage du génome
GWAS : Genome-Wide Association Study
HERV : Human Endogenous RetroVirus

IRM : Imagerie par Résonance Magnétique

LHON : Leber hereditary optic neuropathy

L1 : LINE-1

LINE : Long Interspersed Nuclear Element

LTR : Long Terminal Repeat

MELT : Mobile Element Locator Tool

MISZ : Z-score indiquant l'intolérance du gène aux variations faux-sens

MitoTIP : Mitochondrial tRNA Informatics Predictor

NGSmt : Séquençage ciblé de l'ADNmt

NHEJ : Non-Homologous End Joining

OMIM : Online Mendelian Inheritance in Man

ORF : Phase ouverte de lecture (« Open Reading Frame »)

PCR : Polymerase chain reaction

PCR-RFLP : PCR-Restriction Fragment Length Polymorphism

pLi : Probabilité qu'un gène soit intolérant à une variation perte de fonction

QI : Quotient Intellectuel

rCRS : revised Cambridge Reference Sequence

SAM : Sequence Alignment/Map

SINE : Short Interspersed Nuclear Element

SNP : Single Nucleotide Polymorphism

SNV : Single Nucleotide Variant

SOLiD : Sequencing by Oligonucleotide Ligation and Detection

SR : Lecture « splittée »

STR : Short Tandem Repeat

SV : Variation de Structure (« Structural Variation »)

SVA : SINE-VNTR-Alu

TAD : Topological Association Domain

TSA : Trouble du Spectre de l'Autisme

TSD : Target Site Duplication

UTR : UnTranslated Region

VCF : Variant Caller Format

VNTR : Variable Number of Tandem Repeat

WT : Wild Type

LISTE DES FIGURES

LISTE DES TABLEAUX

TABLE DES MATIÈRES

INTRODUCTION

Une pathologie rare touche moins d'un individu sur 2000. En France, on estime que 3 à 4 millions d'individus seraient touchés par une maladie rare dont la moitié d'enfants. Un tiers des patients n'aurait pas de diagnostic posé, ce qui constitue un enjeu majeur de santé publique (Fondation Maladies Rares). Actuellement, entre 6000 et 8000 maladies rares sont répertoriées. Environ 2000 d'entre elles concernent les anomalies du développement embryonnaire d'origine génétique avec ou sans déficience intellectuelle. On estime que 2 à 4 % des naissances sont touchées. Les anomalies du développement (AD) sont des troubles du développement anatomique et/ou neurologique qui se manifestent au cours de la vie embryonnaire ou fœtale en réponse à des anomalies génétiques, chromosomiques ou à l'action de l'environnement. L'origine génétique de ces anomalies concernerait 80 % des cas (Liu et al., 2019). La transmission de ces affections peut être dominante, récessive ou liée à l'X. La cause génétique d'environ 5700 d'entre elles a été identifiée. Il reste donc des pathologies à élucider.

L'étude de ces maladies rares est nécessaire afin d'avancer dans les connaissances de leurs bases moléculaires. L'objectif est de réduire la période entre l'apparition des symptômes et le diagnostic de la maladie, appelée errance diagnostique, et de permettre une meilleure prise en charge des patients afin d'améliorer leur quotidien. Au niveau familial, un diagnostic posé permet de proposer un conseil génétique et d'anticiper les récidives de pathologies souvent graves. L'essor du séquençage à haut débit en routine a permis de progresser de manière significative dans le diagnostic des pathologies génétiques rares. Néanmoins, plus de 60 % des cas ne sont toujours pas résolus.

Ce travail de thèse a donc porté sur l'amélioration de l'analyse de données de séquençage à haut débit d'exome (ES) appliquée à l'étude de pathologies génétiques rares, en particulier avec AD.

GÉNÉRALITÉS

I- LE GÉNOME HUMAIN

I.1- L'ADN nucléaire

I.1.1- Description et structure

L'ADN nucléaire est localisé au sein du noyau des cellules eucaryotes. C'est un ADN bicaténaire linéaire organisé en 23 paires de chromosomes homologues pour une taille totale d'environ 3,2 milliards de bases. Chaque paire de chromosomes a une origine biparentale. On estime que 1,5-2 % de cet ADN est constitué de régions codant pour des protéines (**Fig. 1**) avec une estimation de 20 000 gènes (Yates et al., 2017).

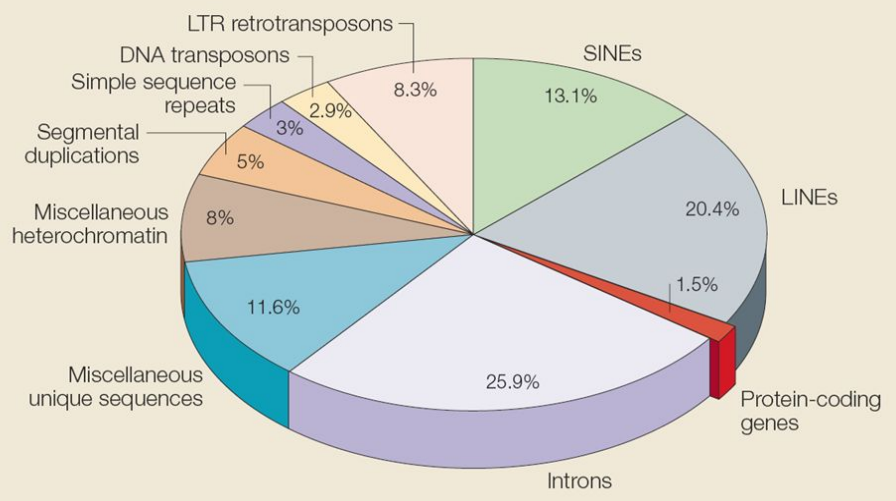


Figure 1 : Composition du génome humain
Plus de 50 % de l'ADN est constitué de séquences répétées (éléments transposables, répétitions, duplications). Les introns et les séquences régulatrices représentent presque 26 %. Les régions codantes sont minoritaires avec 1,5 % du génome. Les pseudogènes, appartiennent à la catégories des séquences uniques qui constituent 11,6 % du génome (d'après (Gregory, 2005)).

La version actuelle (en date du 28 février 2019) du génome de référence est la GRCh38.p13 maintenue par le Genome Reference Consortium ou hg38 selon l'Université de Californie (UCSC). Le taux de mutation de l'ADN nucléaire est estimé à environ 10^{-8} par site par génération (1000 Genomes Project Consortium, 2010).

I.1.2- Variations génomiques (SNV, indel, CNV et variations de structures)

Les variations ponctuelles ou SNV (« Single-Nucleotide Variant ») correspondent à la modification d'un nucléotide par rapport à la séquence de référence. Celles qui ont une fréquence supérieure à 1 % dans la population sont considérées comme des polymorphismes ou SNP (« Single-Nucleotide Polymorphism ») (1000 Genomes Project Consortium, 2010). A l'inverse, un SNV de fréquence inférieure à 1 % est qualifié de rare. Ces variations peuvent être silencieuses (pas de modification de l'acide aminé), faux-sens (changement de l'acide aminé) ou non-sens (changement en codon stop). L'apparition d'un codon stop peut également être provoqué par une modification ponctuelle d'un site d'épissage ou par la modification du cadre de lecture (insertion ou délétion d'un nucléotide). Les insertions ou délétions de moins de 10 000 nucléotides sont appelées « indels ». Elles représenteraient entre 16 et 25 % des polymorphismes (Mills et al., 2006). Néanmoins une indel dont la longueur n'est pas multiple de 3 bases provoque un décalage du cadre de lecture qui peut être à l'origine de pathologie (Mullaney et al., 2010).

Les anomalies de structure (SV pour « Structural Variation ») sont le résultat de cassures chromosomiques suivies d'une réparation anormale. Lorsqu'il n'y a ni perte ni gain de matériel génétique, les anomalies sont dites équilibrées. Ces dernières sont majoritairement des translocations équilibrées, des insertions ou des inversions. Les porteurs peuvent être sains mais ces anomalies sont susceptibles de provoquer l'apparition de pathologies génétiques pour la descendance si un point de cassure implique un gène morbide.

Les variations de structure déséquilibrées sont moins fréquentes au sein de la population et sont à l'origine d'une perte ou d'un gain de matériel génétique. En premier lieu se trouvent les variations du nombre de copies ou CNV (« Copy Number Variation »). Il s'agit de délétions ou de duplications de régions chromosomiques. Un grand nombre de CNV ont été décrits en pathologies humaines (Zhang et al., 2009) comme le syndrome de Prader-Willi (OMIM 176270) causé par une délétion dans la région 15q11-13 ou le syndrome de la duplication de la région 22q11 (OMIM 608363).

Dans les SV peut également être inclus l'insertion d'éléments mobiles (cf section I.2). Enfin il existe des anomalies complexes qui impliquent plus de 2 chromosomes.

I.1.3- Hérédité mendélienne

L'ADN nucléaire est d'origine biparentale. L'hérédité des pathologies nucléaires est donc de 3 types : autosomique dominante, autosomique récessive ou transmission liée à l'X (**Fig. 2**).

Les modes de transmission autosomique dominant ou récessif concernent les variations au sein des autosomes. Ils touchent donc indifféremment les hommes et les femmes. Dans une pathologie de transmission autosomique dominante, l'allèle muté est présent à l'état hétérozygote et s'exprime au détriment de l'allèle non muté. Dans une pathologie de transmission autosomique récessive, la présence de deux allèles mutés au sein du même gène est nécessaire. Les malades sont soit homozygotes pour une même variation soit hétérozygotes composites avec une variation différente sur chacun des homologues. Les porteurs hétérozygotes ne sont donc pas atteints par la maladie. Ainsi, un patient homozygote ou hétérozygote composite aura des parents hétérozygotes pour le ou les allèles mutés. Les familles consanguines présentent donc un risque plus élevé d'avoir une descendance atteinte d'une pathologie récessive qu'un couple non apparenté.

L'hérédité liée à l'X peut être dominante ou récessive. Dans l'hérédité dominante liée à l'X, les hommes hémizygotes et les femmes hétérozygotes sont touchés. Il peut arriver que la pathologie n'atteigne que les femmes si elle s'avère létale chez les hommes. En cas de récessivité, seuls les garçons porteurs de la variation sont atteints. Les femmes sont conductrices et n'expriment pas la maladie.

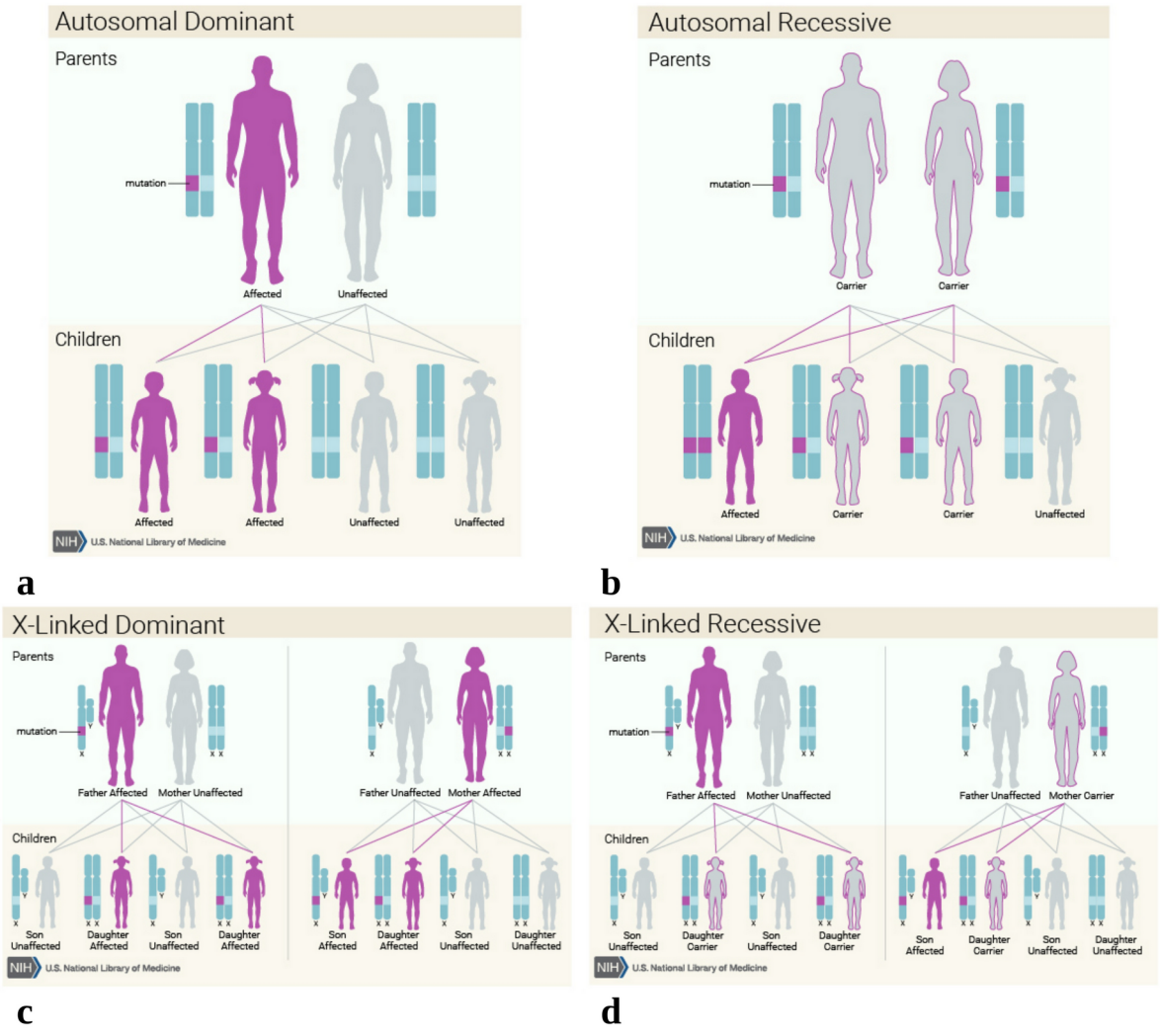


Figure 2 : Hérédité mendélienne de l'ADN nucléaire

a) Hérédité autosomique dominante. Seul un allèle muté suffit pour exprimer la maladie aussi bien chez les hommes que chez les femmes. Les descendants ont 50 % de risque de recevoir l'allèle muté et donc de présenter eux-aussi la maladie. Il s'agit également du mode d'hérédité de toute variation de novo hétérozygote causale présente au sein d'un autosome. b) Hérédité autosomique récessive. La présence de deux allèles mutés est nécessaire pour exprimer la pathologie quel que soit le sexe de l'individu. Les hétérozygotes sont porteurs sains et ont 50 % de risque de transmettre l'allèle muté. Ainsi un couple hétérozygote a 25 % de risque d'avoir un enfant atteint car homozygote ou hétérozygote composite. c) Hérédité dominante liée à l'X. La variation causale est située sur le chromosome X. La pathologie dominante s'exprime indifféremment chez les hommes et chez les femmes. Un homme atteint ne peut pas transmettre la variation à ses fils mais ses filles hétérozygotes sont nécessairement porteuses et malades. Une femme atteinte a 50 % de risque de transmettre son chromosome X muté à chacun de ses enfants (fille hétérozygote ou garçon hémizygote) qui seront donc touchés. Les autres enfants ne seront donc ni porteurs ni atteints. d) Hérédité récessive liée à l'X. La variation causale est située sur le chromosome X. La pathologie récessive s'exprime chez les hommes hémizygotes. Un homme atteint ne peut pas transmettre la variation à ses fils mais ses filles hétérozygotes sont nécessairement porteuses mais non atteintes : elles sont conductrices. Une femme porteuse n'exprime pas le maladie. Elle n'a pas de filles malades mais 50 % sont porteuses. En revanche, ses fils ont 50 % de risque d'être porteurs hémizygotes donc malades et 50 % de chance de ne pas hériter de l'allèle muté.

Les individus atteints sont représentés en violet, les individus sains en gris et les individus porteurs en gris entourés de violet (d'après (NIH)).

I.2- Particularité des éléments mobiles

I.2.1- Définition et description

Parmi les défis de l'analyse de données d'exome et de génome, se trouve l'identification des éléments mobiles appartenant à la famille des éléments transposables (ET). Ces derniers sont des séquences répétées ayant ou ayant eu la capacité de se déplacer au sein du génome. Ils constituent environ 45 % du génome (**Fig. 3**) et dupliquent leur site cible lors de leur insertion. De plus, ils peuvent s'insérer dans toutes les régions du génome, comme des gènes ou des régions régulatrices, et être responsables d'événements comportant dans le premier cas, une modification du cadre de lecture et dans le second cas, une dérégulation de l'expression génique (Kim et al., 1998). Des méthodes existent pour identifier les rétrotransposons, mais elles ne sont pas classiquement intégrées dans les pipelines d'analyse de données de ES ou de génome.

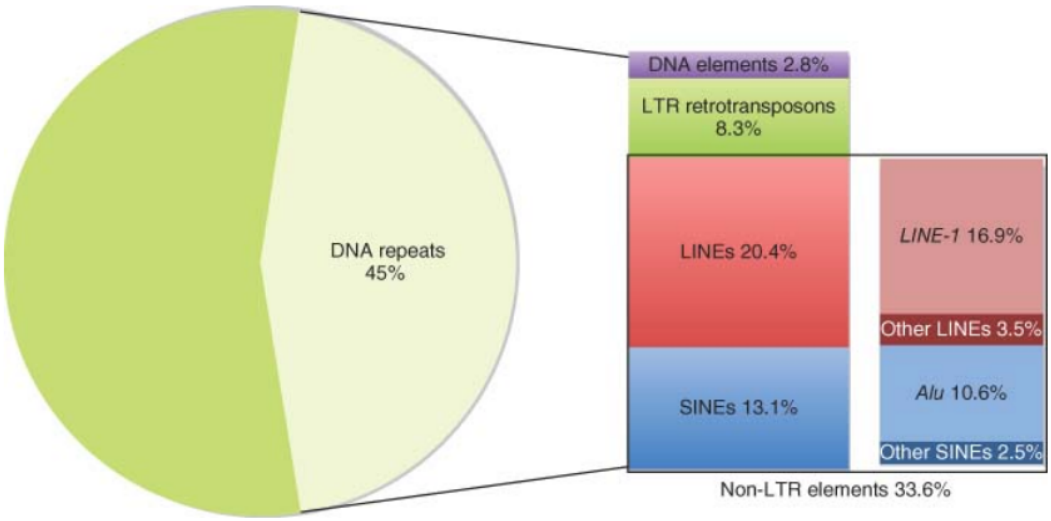


Figure 3 : Diversité des éléments transposables au sein du génome humain
Les éléments mobiles constituent 27,7 % du génome humain avec 16,9 % de L1, 10,6 % d'Alu et 0,2 % de SVA (d'après (Chenais, 2015)).

Les ET se classent en 2 catégories (Wicker et al., 2007) en fonction de leur mécanisme d'insertion dans le génome (transposition) :

- les rétrotransposons ou éléments de classe I, sont des éléments à ARN (42,2 % du génome).

- les transposons ou éléments de classe II, sont des éléments à ADN (2,8 % du génome).

Chez l'être humain, les éléments mobiles (EM) appartiennent à la classe I. Cette dernière est composée de deux sous-groupes, les rétrotransposons à LTR (« Long Terminal Repeat ») et les rétrotransposons non-LTR. Seuls les derniers sont mobiles. Ils se multiplient dans le génome et sont ainsi à l'origine de séquences répétées dispersées (Wicker et al., 2007). Ces éléments s'amplifient par un processus de « copier-coller » semblable à celui utilisé par les rétrovirus ; et qui fait intervenir la reverse-transcription d'un ARN. Les autres catégories d'ET (classe II et rétrotransposons à LTR) ne sont plus actives chez l'humain.

I.2.2- Structure des rétrotransposons non-LTR

Ces éléments mobiles représentent environ 33,6 % du génome humain (**Fig. 3**). Ils ne possèdent pas de LTR mais une région riche en A à leur extrémité 3' ou queue polyA. De part et d'autre de ces éléments se trouvent les TSD (« Target Site Duplications »), constituées de répétitions de fragments de 2 à 20 pb, qui correspondent à des duplications des sites d'insertion autour du nouvel EM transposé (Goerner-Potvin and Bourque, 2018).
Les rétrotransposons non-LTR se divisent en 3 catégories : les LINE (« Long Interspersed Nuclear Elements »), les SINE (« Short Nuclear Interspersed Elements ») et les SVA (SINE-VNTR-Alu).

Les LINE (21 % du génome humain), d'environ 5-8 kd, possèdent deux ORF (*gag* et *pol*), mais contrairement aux rétrotransposons à LTR, ils ne codent pas pour une intégrase d'où l'existence d'un mécanisme de rétrotransposition différent de ceux-ci (**Fig. 4**). L'ORF1 code pour une protéine qui forme avec l'ARN un complexe ribonucléoprotéique (RNP). L'ORF2 code pour une protéine aux activités endonucléase et reverse transcriptase qui vont permettre : la coupure double-brin de la région cible de l'ADN, la reverse transcription de l'ARN et l'intégration de l'ADNc au niveau de la coupure. Ils possèdent en 5' un promoteur pour l'ARN polymérase II ; et en 3' une queue polyA. Chez l'humain, seul l'élément LINE-1 (L1) est encore mobile (International Human Genome Sequencing Consortium, 2001), il représente 17 % du génome avec un nombre de copies estimé à plus de 500 000 (Cordaux and Batzer, 2009).

Les SINE (12 % du génome humain), d'environ 300 pb, ne possèdent pas de région codante (**Fig. 4**). Ils possèdent des régions Box A et B du promoteur de la polymérase III. La majorité des éléments SINE et des éléments mobiles sont des éléments Alu (10 % du génome avec plus d'un million de copies au sein du génome) (Cordaux and Batzer, 2009). Ils dérivent de l'ARN 7SL, composant de la particule de reconnaissance du signal qui intervient dans la traduction des ARNm (Zwieb et al., 2005).

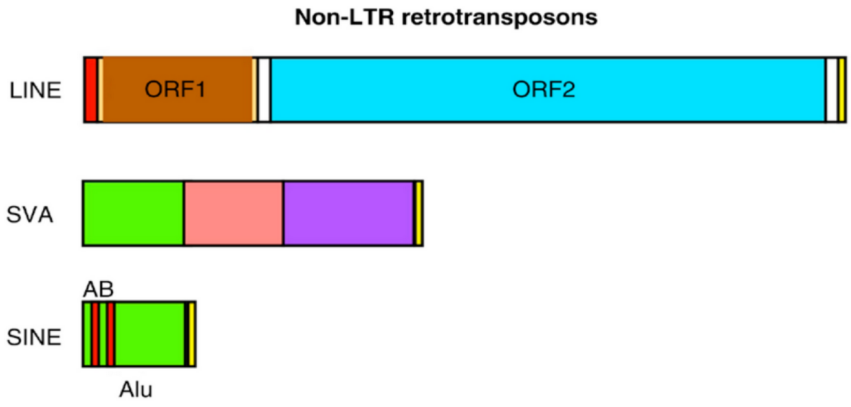


Figure 4 : Structure des éléments mobiles
Les LINE sont transcrits à partir d'un promoteur (en rouge). L'ORF1 (en marron) code pour une protéine de liaison à l'ARN. L'ORF2 (en bleu) code pour une protéine endonucléase (RNase H) et reverse transcriptase. Enfin, les LINE se terminent par une queue polyA (en jaune). Les SVA ne contiennent pas de séquence codante : ce sont des éléments non-autonomes. En 5' se trouvent des répétitions d'hexamères suivies de deux séquences qui dérivent d'Alu en antisens (en vert). Ils possèdent également une région VNTR (en rose), suivie de l'extrémité 3' du rétrovirus HERV-K10 (en violet) et d'une queue polyA (en jaune). Les SINE sont également des éléments non-autonomes. Les éléments Alu dérivent de l'ARN SL (en vert), contiennent des box A et B du promoteur de l'ARN polymérase III (en rouge) et une queue polyA (en jaune) (d'après (Finnegan, 2012)).

Les SVA (0,2 % du génome humain (Hancks and Kazazian, 2010)), d'environ 2 kb, ne possèdent pas de région codante (**Fig. 4**). Ils sont constitués en 5' de répétitions de l'hexamère CCCTCT, suivis d'une région homologue composée de deux fragments homologues à Alu antisens, d'une région VNTR (Variable Number of Tandem Repeat) constituée d'un nombre variable de répétitions en tandem de séquences riches en GC de 35 à 50 pb, d'un élément SINE d'environ 490 pb dérivant de l'extrémité 3' du gène *env* et de l'extrémité 3' LTR du rétrovirus endogène HERV-K10, et d'une queue polyA. Environ 3000 copies de SVA seraient présentes au sein du génome humain (Cordaux and Batzer, 2009).

I.2.3- Mécanismes de rétrotransposition des rétrotransposons non-LTR

Les L1 sont les seuls rétrotransposons non-LTR à être des éléments autonomes. La transcription de cette région est régulée par la méthylation du promoteur. Lorsque ce dernier est déméthylé, le L1 est transcrit par une ARN polymérase II en ARN, qui est exporté dans le cytoplasme pour y être traduit. Le complexe RNP, formé par la protéine nouvellement synthétisée et l'ARN, est importé dans le noyau. La nucléase identifie et coupe un site riche en A/T de l'ADN receveur (**Fig. 5**). La queue polyA en fin de l'ARN permet l'initiation de la synthèse du premier brin de la molécule d'ADN (Finnegan, 2012). Le second brin est ensuite synthétisé à partir de l'extrémité 3'OH du brin complémentaire déjà intégré au génome. L'intégration des L1 ne fait donc pas intervenir d'intégrase.

Les SINE sont des éléments non autonomes qui utilisent la machinerie d'intégration des L1 (Finnegan, 2012) car ils n'ont pas de séquence codante. L'extrémité 3' est similaire à celle des éléments L1 et contient donc un promoteur interne de l'ARN polymérase III qui permet la future transcription en ARN de ces éléments transposables nouvellement intégrés.

Les SVA sont également des éléments non autonomes qui utilisent la machinerie d'intégration des L1 faisant appel à l'ARN polymérase II (Wang et al., 2005; Raiz et al., 2012). En effet l'ajout d'une queue polyA aux ARN transcrits à partir d'un SVA tronqué en 3' et la présence d'un résidu G en 5' (coiffe) sont des éléments mettant en évidence une activité de l'ARN polymérase II (Raiz et al., 2012).

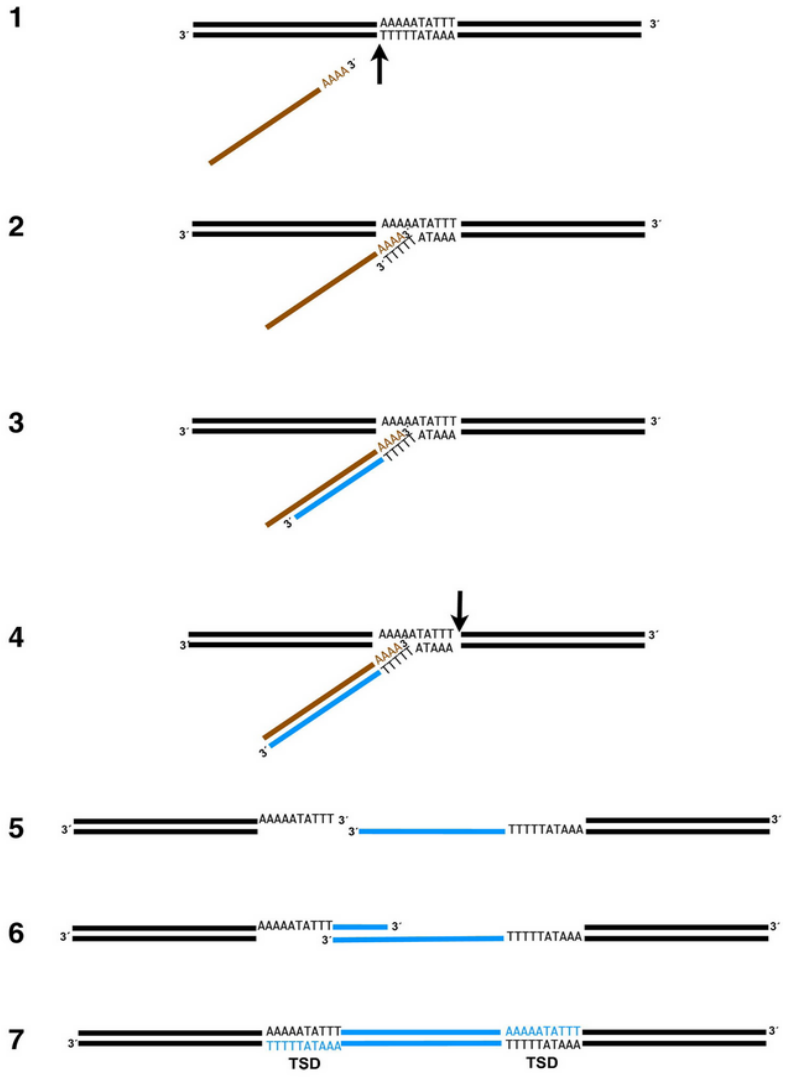


Figure 5 : Mécanisme de rétrotransposition de l'élément L1
1) Identification d'une région riche en A/T et cassure du brin anti-sens par l'endonucléase. 2) Hybridation de la queue polyA. 3) Synthèse du premier brin d'ADN par la reverse transcriptase. 4) Cassure du brin sens par l'endonucléase. 5) Dégradation de l'ARN par la RNase H. 6) Initiation de la synthèse du brin complémentaire en 3' du brin nouvellement synthétisé. 7) Réparation de la cassure DB par les enzymes cellulaires. TSD : Terminal Side Duplication (d'après (Finnegan, 2012)).

I.2.4- Impacts des EM sur le génome

L'impact des EM sur le génome a lieu à plusieurs niveaux (Cordaux and Batzer, 2009). Le phénomène de rétrotransposition peut créer une instabilité génomique locale ou à plus grande échelle (réarrangements génomiques) à l'origine de pathologies, participer à la création de nouveaux gènes ou impacter l'expression génique.

I.2.4.1- Instabilité locale

À l'échelle locale, un des impacts sur le génome le plus facilement détectable est l'insertion *de novo* d'un rétrotransposon L1, Alu ou SVA au sein de séquences codantes ou de promoteurs (**Fig. 6a**). Ainsi, on considère qu'environ 0,3 % des variations pathogènes du génome humain sont dues à l'insertion *de novo* d'un EM. Par exemple, une des causes de l'hémophilie B est l'insertion d'un élément Alu dans le gène *F9* (Deininger and Batzer, 1999).

De plus, de par son activité, l'endonucléase codée par l'élément L1 provoque des cassures doubles-brins au sein de l'ADN (**Fig. 6b**). Il a été montré que le nombre de ces cassures est plus élevé que le nombre d'insertions réelles de l'élément L1 (Gasior et al., 2006). L'instabilité génomique qui découle des cassures doubles-brins aurait donc pour origine l'intervention d'un élément mobile. Mais il a également été mis en évidence l'intervention des éléments L1 et Alu dans la réparation de ces cassures (Morrish et al., 2002). Ces éléments mobiles ont la capacité de s'insérer au niveau de ces cassures préexistantes via la reverse transcriptase. Cette rétrotransposition est donc indépendante de l'endonucléase. Elle a été observée notamment au sein des cellules déficientes pour le système de réparation par jonction d'extrémités non homologues ou NHEJ (« Non-Homologous End Joining »). Les éléments mobiles ainsi insérés ne possèdent pas la structure habituelle des éléments mobiles : ils sont souvent tronqués à l'extrémité 3' et ne possèdent pas de TSD. Ils représenteraient 0,5 à 0,7 % des éléments Alu et L1 insérés (Sen et al., 2007 ; Srikanta et al., 2009). Ce mécanisme participerait donc au maintien de l'intégrité du génome via un système de réparation des cassures doubles-brins par insertion d'élément mobile.

Une des caractéristiques des rétrotransposons non-LTR est qu'ils possèdent des régions répétées pouvant être à l'origine de microsatellites (**Fig. 6c**). Le cas le plus étudié est celui de

l'élément Alu avec sa séquence riche en A et sa queue polyA. L'ataxie de Friedreich (OMIM 229300) est l'un des exemples de l'instabilité génomique créée par les microsatellites. Ces derniers augmentent le risque de substitutions nucléotidiques et de glissement du brin d'ADN répliqué, à l'origine de l'expansion de triplets GAA conduisant à la maladie.

Enfin le phénomène de conversion génique participe à l'instabilité génomique au niveau local **(Fig. 6d**). Ce processus, qui concerne les éléments Alu, consiste en un transfert unidirectionnel d'information génétique (séquence) entre une séquence 'donneuse' et une séquence réceptrice à forte homologie et contenant un point de cassure double-brin (Chen et al., 2007). Cette altération de la séquence peut conduire à une inactivation génique. Ce processus participe à l'évolution des espèces. Le gène CMP-N-acetylneuraminic acid hydroxylase a, par exemple, été inactivé chez l'espèce humaine par remplacement de la séquence AluSq qui s'y trouvait par un élément AluY (Hayakawa et al., 2001). Cette conversion génique a provoqué la délétion de 92 pb inactivant ainsi le gène.

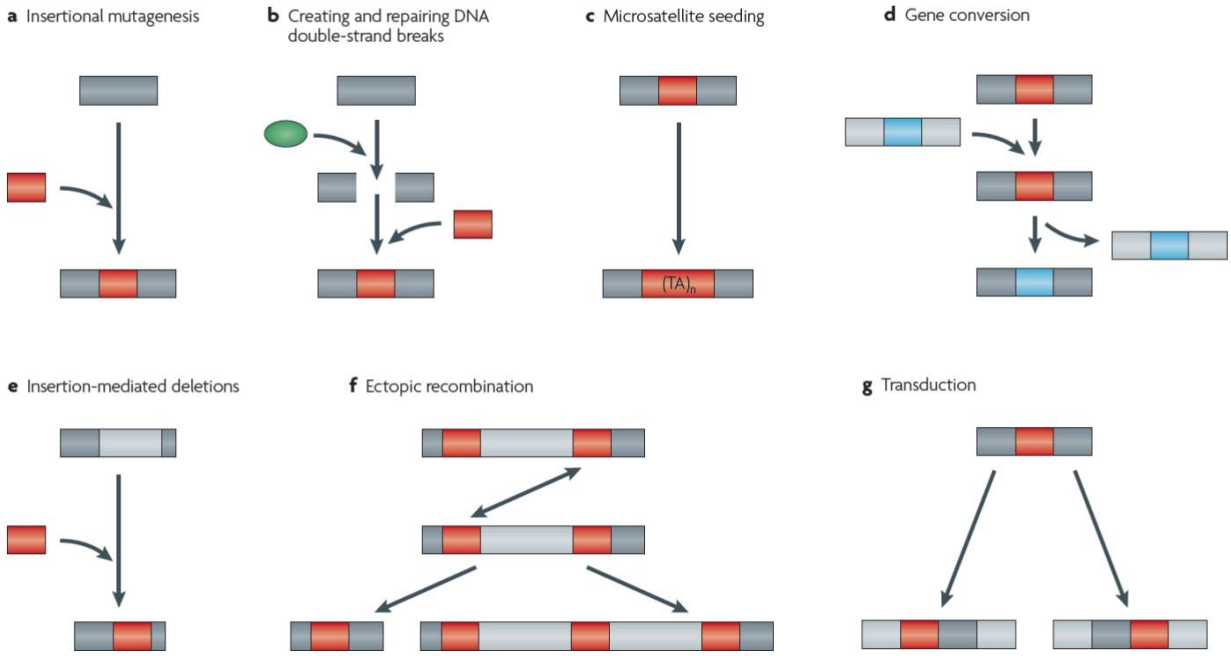


Figure 6 : Impacts locales ou structurels des éléments mobiles sur le génome
a) Insertion de novo d'un élément mobile (en rouge) au sein d'une région génique (en gris foncé) impactant la séquence codante. b) Cassure double-brin par l'endonucléase de l'élément L1 (en vert) et réparation par l'insertion d'un élément mobile (en rouge). c) Accumulation de régions répétées d'éléments mobiles (en rouge) à l'origine d'un microsatellite. d) Conversion génique par remplacement partiel ou total d'un élément Alu (en rouge) par un autre élément Alu (en bleu) à forte homologie de séquence. Les variations au sein de la séquence génique initiale peuvent conduire à l'inactivation du gène. e) Délétion de la région adjacente (en gris clair) au point de cassure au cours de l'insertion d'un élément L1 ou Alu (en rouge). f) Recombinaison ectopique des éléments L1 ou Alu (en rouge) à l'origine de variations de structure. g) Transductions des régions géniques flanquantes en 5' ou en 3' (en gris clair) de l'élément mobile L1 ou SVA (en rouge) lors de la rétrotransposition de ce dernier (d'après (Cordaux and Batzer, 2009)).

I.2.4.2- Réarrangements génomiques

L'impact des EM n'est pas uniquement locale mais peut également être à l'origine de réarrangements génomiques. Par exemple, l'insertion des éléments mobiles peut provoquer une délétion de la région génomique cible (**Fig. 6e**). Ce processus participe à l'évolution des espèces avec l'exemple de la disparition d'un gène de récepteur olfactif des génomes humain et de chimpanzé par insertion d'un élément L1 (Sen et al., 2007).

Mais la présence d'éléments mobiles au sein du génome peut être également à l'origine de la création de variations de structure hors processus de rétrotransposition. En effet, les éléments L1 et Alu sont très nombreux dans le génome et peuvent subir une recombinaison homologue non allélique (recombinaison ectopique) conduisant à des délétions, duplications ou

inversions (**Fig. 6f**). Plusieurs pathologies ont été décrites comme ayant pour origine une recombinaison d'éléments Alu (Deininger and Batzer, 1999) ou L1 (Ostertag and Kazazian Jr, 2001).

Lors de la rétrotransposition, les EM peuvent également emporter une partie des régions flanquantes en 5' ou en 3' (**Fig. 6g**). Ce phénomène est appelé transduction 5' ou 3'. La transduction 5' a lieu lorsqu'il y a transcription de la séquence de l'élément mobile mais également de la région en 5' grâce à un promoteur situé au sein de cette dernière. Ce mécanisme concerne majoritairement les éléments SVA. La séquence génomique ainsi transcrite sera intégrée à un autre endroit du génome lors de la rétrotransposition. Il peut y avoir ainsi création de nouvelles combinaisons de régions géniques et régulatrices, processus participant à la plasticité du génome (Damert et al., 2009). La transduction 3' est, quant à elle, plus fréquente : on estime qu'environ 10 % des SVA (Wang et al., 2005) et 9 % des L1 (Szak et al., 2003) ont transduit leur région en 3' lors de leur rétrotransposition. Ce phénomène a lieu lorsque l'ARN polymérase II passe outre le signal de polyadénylation de l'élément mobile et en utilise un autre situé plus en 3'. Cette transduction 3' propage des séquences géniques, exoniques et de régulation au sein de nouvelles régions du génome (Tubio et al., 2014) et a participé au façonnage du génome. Ces transductions 5' et 3' participent à l'évolution des espèces notamment par le processus d'« exon-shuffling », littéralement le « brassage d'exons », à l'origine de la création de nouveaux gènes.

L'innovation génétique, ou la création de nouveaux gènes est en effet un autre des impacts de la rétrotransposition sur le génome, par transduction de régions codantes à l'origine de la copie et du remaniement de séquences exoniques. Ce processus participe à l'évolution du génome humain avec l'exemple de la famille de gènes *AMAC1* (Xing et al., 2006). Ce groupe est constitué de 4 gènes ayant pour origine la transduction 3' d'un élément SVA. Un élément SVA s'est inséré en 5' du gène *SLC35G6* (anciennement *AMAC1L3*), localisé en 17p13.1. S'en est suivie une rétrotransposition de cet élément mobile qui a emporté la séquence du gène ainsi que ses régions régulatrices et promotrices. Trois nouveaux gènes issus de gène ont ainsi été créés : les gènes *CCL18* (OMIM 603757, anciennement *AMAC1*) en 17q12, *SLC35G4* (*AMAC1L1*) en 18p11.21 et *SLC35G5* (OMIM 615199, anciennement *AMAC1L2*) en 8p23.1.

A l'inverse il existe des rétrotranspositions qui copient uniquement les séquences géniques sans l'élément mobile impliqué (Esnault et al., 2000). La rétrotransposition de ces gènes ne concernant pas les régions régulatrices, les séquences géniques dupliquées ne pourront être fonctionnelles qu'à la condition d'acquérir une nouvelle région promotrice. Les copies de gènes non fonctionnelles sont nommées rétro-pseudogènes et sont importantes dans l'évolution du génome humain et l'apparition de nouveaux gènes comme par exemple *TRAPPC2B* (anciennement *MIP-2A* ou *SEDLP1*) localisé en 19p13.43 et issu de la copie du gène *TRAPPC2* (OMIM 300202) localisé en Xp22.2. La copie de ce dernier s'est intégrée au sein du gène *ZNF547* dont il utilise le premier exon comme région 5'-UTR ainsi que les régions régulatrices (Ghosh et al., 2001 ; Vinckenbosch et al., 2006).

Un élément mobile peut également s'intégrer au sein d'une région génique et « s'exoniser », c'est-à-dire devenir un exon à part entière du gène considéré (**Fig. 7a**). De plus la séquence d'un grand nombre d'éléments transposables contient des sites donneurs et accepteurs d'épissage. L'exonisation peut donc conduire à l'apparition d'un épissage alternatif au sein du gène. On estime qu'environ 0,1 % des séquences protéiques ont pour origine la séquence d'anciens éléments transposables exonisés au cours de l'évolution (Gotea and Makalowski, 2006). L'exonisation d'éléments récents comme les Alu ou les L1 semblent au contraire avoir un impact négatif sur la fonction des protéines au sein desquelles ils sont insérés en cas d'expression constitutive. L'une des origines du syndrome d'Alport, par exemple, est l'exonisation d'un élément Alu dans le gène *COL4A5* (Nozu et al., 2014). Il existe donc un épissage alternatif et une sélection négative de ces exons exprimés constitutivement (Sorek et al., 2002).

I.2.4.3- Modification de l'expression génique

La rétrotransposition d'éléments mobiles n'a pas uniquement un impact au niveau de la structure du génome ou des gènes mais joue également un rôle dans la modification de l'expression génique aussi bien au niveau transcriptionnel que post-transcriptionnel.

Comme vu précédemment, l'insertion d'un élément mobile au sein d'une région génique peut conduire à une modification de l'épissage, à l'apparition d'un épissage alternatif et à l'exonisation de l'élément mobile inséré (**Fig. 7a**). A l'image des éléments Alu, l'insertion d'un

élément L1 au sein de régions géniques peut être à l'origine de pathologies. Ainsi en 1999, Kondo-lida et son équipe ont décrit un patient atteint d'une dystrophie musculaire congénitale type Fukuyama (OMIM 253800) causée par l'insertion d'un élément L1 au sein d'un intron du gène *FKTN* (OMIM 607440) (Kondo-Iida et al., 1999). Cette insertion a été décrite comme impactant l'épissage en provoquant des sauts d'exons. De même, en 2000, Meischl et son équipe ont décrit un patient atteint de granulomatose chronique (OMIM 306400) causée par l'exonisation d'un élément L1 inséré au sein de l'intron 5 du gène *CYBB* (OMIM 300481) (Meischl et al., 2000).

Une autre façon d'influer sur l'expression génique est d'agir au niveau du déroulement de la transcription (**Fig. 7b**). L'insertion sens ou anti-sens d'un élément L1 au sein d'une région intronique peut avoir un impact sur l'étape d'élongation (Han et al., 2004). En cas d'insertion sens, l'ARN polymérase rencontre des difficultés à transcrire l'intégralité de la longue séquence de l'élément L1 ajoutée à celle du gène. Cela conduit à une diminution de la quantité d'ARN pré-messager du gène considéré. Le signal de polyadénylation présent au sein de la séquence de l'élément mobile peut également conduire à une polyadénylation prématurée si l'ARN polymérase a réussi transcrire l'intégralité de la séquence de l'élément L1. C'est également cette polyadénylation prématurée qui conduit une inhibition de la transcription complète du gène en cas d'insertion anti-sens de l'élément mobile. Les ARN ainsi produits sont donc tronqués, impactant qualitativement l'expression génique.

Les éléments Alu possèdent au sein de leur séquence des régions Box A et B du promoteur de la polymérase III, enzyme transcrivant les Alu (**Fig. 7c**). De plus, il existe des sites de régulation de l'ARN polymérase II entre ces Box A et B (Shankar et al., 2004). Il a été montré que les éléments Alu ne sont pas distribués aléatoirement au sein du génome mais préférentiellement à proximité des gènes impliqués dans la signalisation, le métabolisme et le transport protéique (Shankar et al., 2004). Ils auraient permis d'apporter des sites de régulation au sein des régions promotrices notamment de gènes du développement (Polak and Domany, 2006). Les éléments Alu ont donc un rôle dans la mise en place de blocs de régulation génique.

La séquence des éléments L1 comporte également dans sa région 5'-UTR un promoteur canonique sens ainsi qu'un promoteur anti-sens ASP (« AntiSens Promoter ») (Faulkner et al., 2009) (**Fig. 7d**). Le premier peut initier la transcription sens des régions en 3' de l'élément diminuant ainsi l'activité du promoteur du gène situé en 3'. Le second est à

l'origine de la transcription anti-sens de la région en 5' de l'élément mobile. Ces deux types de promoteurs alternatifs jouent donc un rôle dans la régulation de l'expression des gènes aux alentours. Ainsi 40 gènes posséderaient des promoteurs alternatifs provenant d'un élément mobile (Speek, 2001). De plus, une quinzaine de gènes ont également été décrits comme ayant été divisés par l'insertion d'un élément mobile (Wheelan et al., 2005). Ce processus peut être à l'origine de la création de nouveaux gènes par fission. En effet, en cas d'insertion d'un élément L1 anti-sens dans une région intronique, le gène concerné a 2 transcrits. Le premier contient les exons en 5' de la région d'insertion et se termine par le signal de polyadénylation de l'élément L1. Le second provient du promoteur ASP et comporte les exons de la région en 3' de l'insertion.

La régulation de l'expression génique passe également par l'édition des ARN des gènes situés à proximité du lieu de rétrotransposition. Un des exemples est la transformation d'adénosines en inosines par déamination (**Fig. 7e**). Cette modification est effectuée sur les ARN pré-messagers doubles-brins par l'enzyme ADAR (« Adenosine Deaminase Acting on RNA ») et permet d'augmenter la diversité des isoformes d'ARN et de protéines. Chez les mammifères il s'agit de la forme d'édition d'ARN la plus fréquente (Slotkin and Nishikura, 2013). L'inosine agissant comme la guanosine, le codon est altéré et l'acide aminé peut être modifié en conséquence. De même, l'inosine préfère se lier à une cytosine ce qui peut affecter la structure secondaire de l'ARN. Cette édition peut également conduire à l'introduction ou l'élimination de sites d'épissage ayant ainsi un impact sur l'exonisation des éléments Alu insérés dans le gène concerné. Enfin, cette modification de nucléotide conduit à l'altération de l'appariement entre 2 ARN ou d'un ARN sur lui-même. La présence de 2 éléments Alu insérés en sens inverse permet la formation de la structure double-brin nécessaire à la protéine ADAR. Il est d'ailleurs estimé que 90 % des éditions A-I ont lieu au sein des séquences Alu des ARNm. S'il s'agit d'éléments Alu insérés dans la région 3'-UTR d'un ARNm cette édition peut conduire à la rétention de cet ARNm au sein du noyau par le complexe p54nrb (Chen et al., 2008). Il y a donc extinction de l'expression du gène concerné.

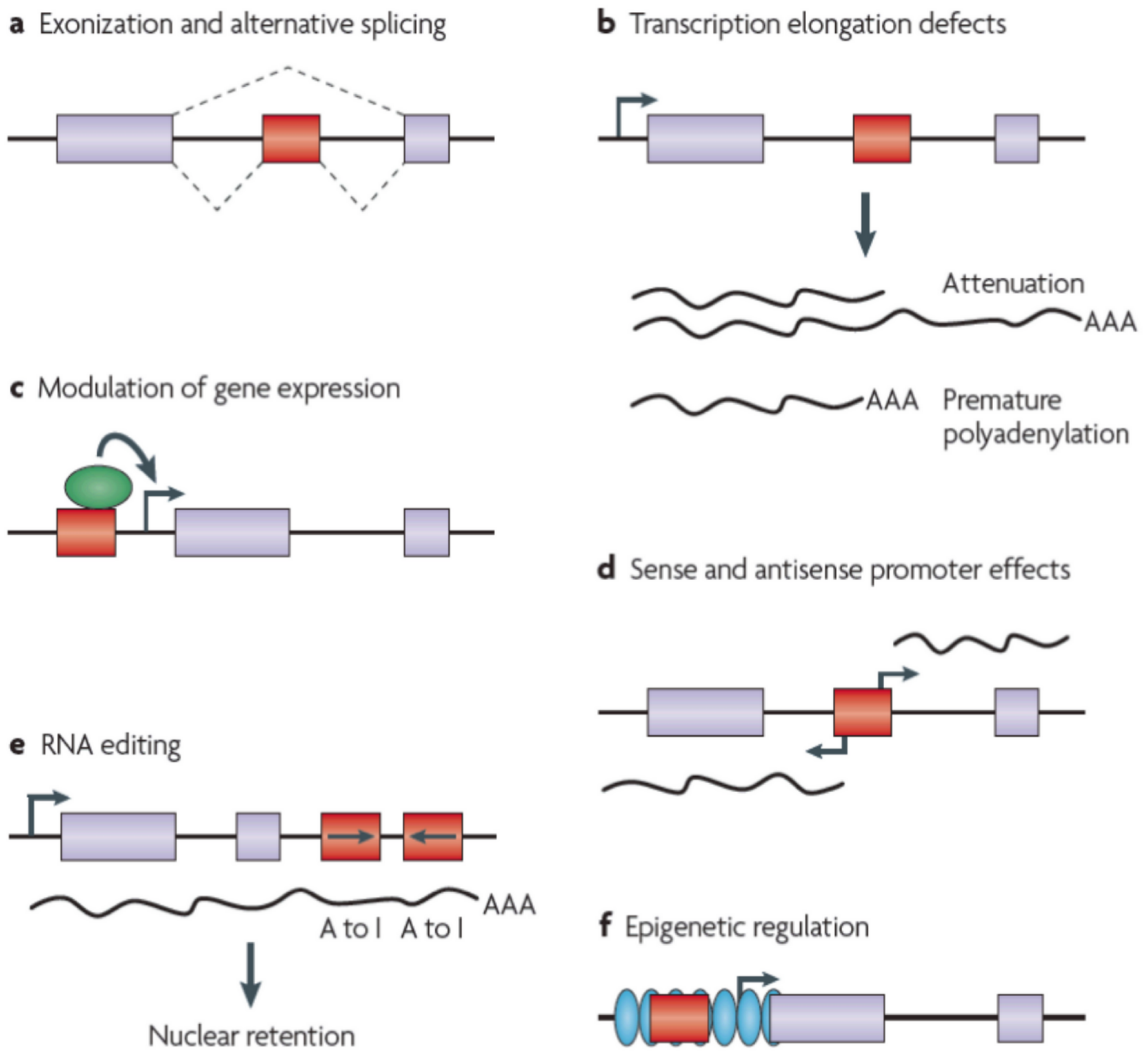


Figure 7 : Impacts des éléments mobiles sur l'expression génique
a) Exonisation et épissage alternatif d'un élément mobile (en rouge) contentant des sites donneurs et accepteurs d'épissage. b) Perturbation de l'élongation par insertion d'un élément L1 (en rouge). Une insertion sens conduit à une diminution du nombre d'ARNm par atténuation de l'activité de l'ARN polymérase. Une insertion anti-sens provoque la polyadénylation prématurée du transcrit. c) Régulation de l'expression génique par fixation de facteurs de transcription (en vert) sur les sites de liaison présents au sein de la séquence de l'élément mobile (en rouge) en 5' du gène. d) Transcription à partir des promoteurs sens et anti-sens (flèches) de l'élément L1 (en rouge). Deux transcrits différents sont synthétisés. Ce processus peut conduire à la création de deux nouveaux gènes. e) Édition de l'ARN par transformation d'adénosines en inosine par l'enzyme ADAR. Les éléments mobiles insérés en sens inverse (en rouge) favorisent la structure secondaire de l'ARN nécessaire au fonctionnement de l'enzyme. Les ARNm peuvent se retrouver bloqués au sein du noyau. f) Régulation épigénétique par méthylation (en bleu) des îlots CpG présents au sein des éléments mobiles (en rouge). La propagation de la méthylation conduit à l'inactivation du gène en 3' de l'élément mobile (d'après (Cordaux and Batzer, 2009)).

Enfin la modulation de l'expression génique peut intervenir au niveau épigénétique par la méthylation de l'ADN au niveau des îlots CpG. Celui présent au sein du promoteur de l'élément L1 peut également être méthylé conduisant à une inactivation de cet élément mobile par la cellule. De plus, les éléments Alu et SVA contiennent de nombreux îlots CpG : on estime qu'un tiers des îlots CpG du génome humain sont présents au sein des séquences Alu. La méthylation d'éléments L1, Alu ou SVA à proximité de gènes peut conduire à l'inactivation de ces derniers par propagation de la méthylation (**Fig. 7f**). C'est l'une des raisons qui expliquerait la faible proportion d'éléments Alu au sein des régions soumises à empreinte (Greally, 2002). Cette propagation de la méthylation apporterait également un élément de réponse quant à l'enrichissement en éléments L1 du chromosome X (Bailey et al., 2000). En effet, la méthylation des L1 permettrait de faciliter l'inactivation d'un des 2 chromosomes X chez la femme.

I.2.5- Éléments mobiles et pathologies

Il a été mis en évidence que les éléments mobiles ont joué un rôle dans l'évolution du génome humain, par exemple en créant ou en supprimant des gènes. De plus, le taux de rétrotransposition varie en fonction de l'élément considéré. Pour les éléments Alu, les plus nombreux au sein du génome, l'estimation est d'une nouvelle insertion pour 20 naissances. Viennent ensuite les éléments L1 pour lesquels les estimations varient entre 1/20 et 1/200 naissances. Enfin pour les SVA, les éléments les moins nombreux, l'estimation est de 1/900 naissances vivantes (Beck et al., 2011). Néanmoins, les impacts de ces éléments au niveau de la structure et/ou de l'expression génique peuvent être à l'origine de pathologies génétiques ou de cancers. Ainsi, 0,3 % des variations pathogènes du génome humain seraient dues à l'insertion *de novo* d'un élément mobile (Cordaux and Batzer, 2009). Le premier cas décrit d'insertion *de novo* pathogène d'un élément mobile date de 1988. Il s'agit d'un élément L1 inséré au sein du gène *F8* codant pour le facteur VIII chez un patient atteint d'hémophilie A (Kazazian et al., 1988). Depuis, de nouveaux cas d'hémophilies, de pathologies génétiques ou de cancers causés par des éléments mobiles ont été rapportés. En 2016, on dénombre ainsi 119 cas décrits dont 76 impliquant un élément L1, 30 un élément Alu et 13 un élément SVA (Hancks and Kazazian, 2016). Il s'agit d'insertions au sein de régions exoniques, introniques, d'épissage ou UTR ;

ainsi que de délétions. L'existence de pathologies causées par la rétrotransposition d'éléments Alu, L1 ou SVA montre l'importance de la détection de ces EM au sein du génome de patients atteints de maladies rares.

I.3- L'ADN mitochondrial

I.3.1- Description et structure

La mitochondrie est un organite cellulaire issue de l'endosymbiose d'une protéobactérie (Margulies, 1981). Sa fonction principale est la production de chaleur et d'énergie pour la cellule sous forme d'ATP, grâce à la respiration cellulaire. Elle intervient également dans d'autres processus cellulaires comme l'apoptose. Cet organite possède, au sein de sa matrice (**Fig. 8**), son propre ADN, vestige de cette endosymbiose.

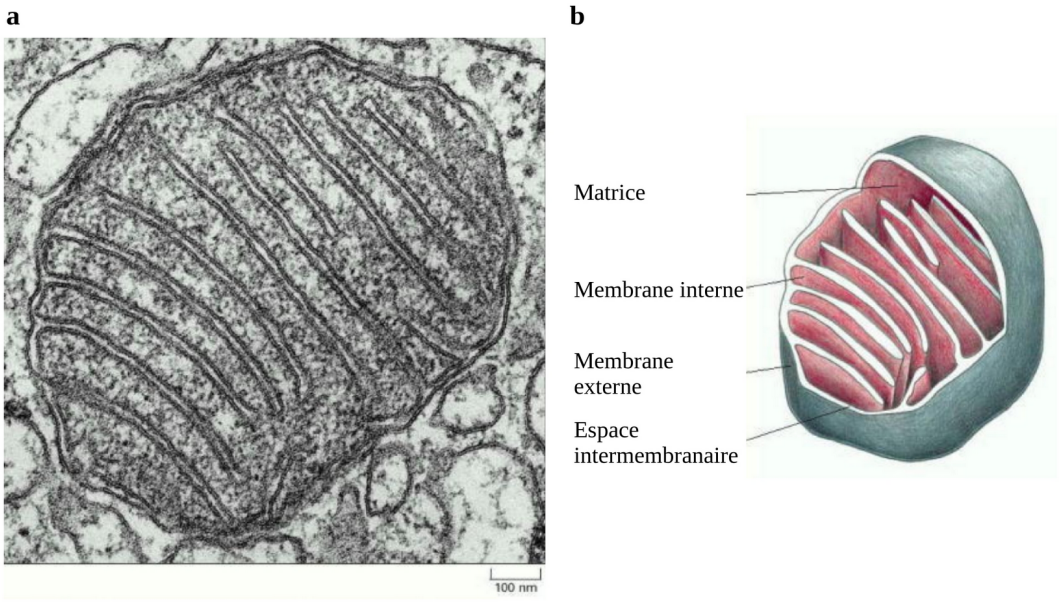


Figure 8 : Structure de la mitochondrie
a) Mitochondrie au microscope électronique. b) Schéma d'une mitochondrie. La matrice est l'espace interne de la mitochondrie et contient plusieurs molécules d'ADNmt. La membrane interne contient les protéines de la chaîne respiratoire mitochondriale. L'espace intermembranaire a une composition proche du cytosol et contient le cytochrome c (d'après (Alberts et al., 2002)).

L’ADN mitochondrial, ou ADNmt, est un ADN double-brin circulaire constitué de 16569 paires de bases ayant son propre code génétique. Sa séquence est numérotée selon la référence *« revised Cambridge Reference Sequence »* (rCRS). Ses deux brins sont de poids moléculaires différents. Le brin L ('Light' = léger) correspond au brin '+' qui sert à l'orientation de la lecture de la base 1 à la base 16569. Le brin H ('Heavy'=lourd) a un poids moléculaire plus important car il est plus riche en guanines. Cette molécule d'ADN double-brin est constituée d'une région régulatrice "D-loop" allant du nucléotide 16076 au nucléotide 576, et de 37 régions géniques. Il n'existe pas de régions introniques et peu de séquences inter-géniques. A l’instar du génome nucléaire, il existe des gènes chevauchants : *MT-ND4*/*MT-ND4L* et *MT-ATP6*/*MT-ATP8*. Ainsi environ 93 % de l'ADNmt est codant (Taylor and Turnbull, 2005). La région régulatrice contient l'origine de réplication (O_H) de cet ADN ainsi que les promoteurs des deux brins (P_H et P_L). Le reste de la séquence nucléotidique contient les 37 gènes mitochondriaux: les gènes *MT-RNR1* et *MT-RNR2* codant pour les 2 ARN ribosomiques 12S et 16S, les gènes *MT-ND1* à *MT-ND6* ainsi que *MT-ND4L* codant pour 7 sous-unités protéiques de la NADH-déshydrogénase (complexe I de la chaîne respiratoire), le gène *MT-CYB* codant pour le cytochrome b de la coenzyme Q-cytochrome c réductase (complexe III), les gènes *MT-CO1* à *MT-CO3* codant pour 3 sous-unités protéiques de la cytochrome c oxydase (complexe IV), les gènes *MT-ATP6* et *MT-ATP8* codant pour deux sous-unités de l'ATP synthase (complexe V) et les 22 gènes correspondant aux 22 ARN de transfert (**Fig. 9**).

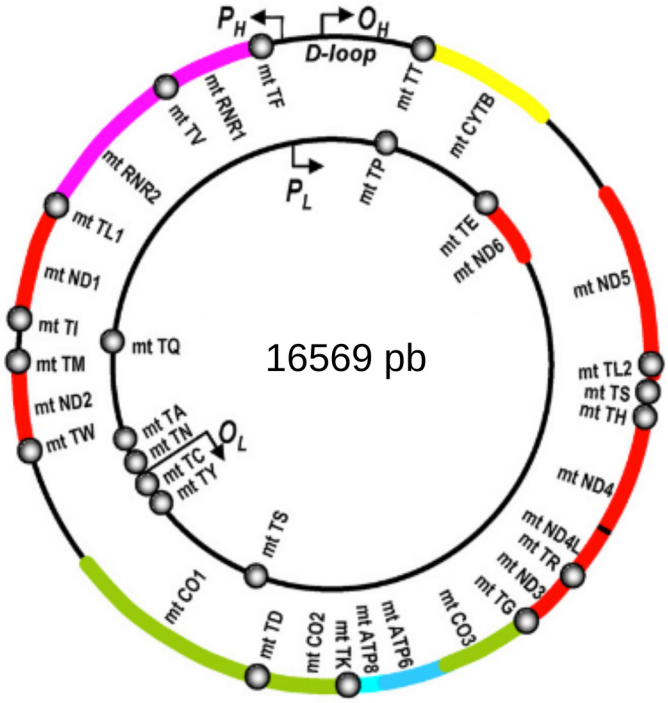


Figure 9 : Représentation de l'ADN mitochondrial
Le brin lourd a son origine de réplication (O_H) au niveau de la boucle D (D-loop). L'origine de réplication du brin léger est indiquée par O_L. Les promoteurs des deux brins sont indiqués par les flèches P_H et P_L. Les différents gènes codant pour les protéines de la chaîne respiratoire et les deux ARNr-mt sont représentés en couleur, ceux codant pour les différents ARNt par des cercles gris (adapté de (Loublier et al., 2009))

Une mitochondrie possède entre 2 et 10 copies d'ADNmt et chaque cellule contient plusieurs mitochondries. L'existence au sein d'une cellule de plusieurs copies d'ADNmt dont la version peut être différente est appelée hétéroplasmie. Ainsi, chez un individu, une variation dans la séquence nucléotidique peut être présente au sein de toutes les molécules d'ADNmt (homoplasmie) ou dans des proportions variables (hétéroplasmie) (**Fig. 10**). En effet, la division mitochondriale, par fission (Okamoto and Shaw, 2005), est indépendante de la division cellulaire. Ainsi, une cellule mère peut engendrer deux cellules filles ayant un contenu en ADN mitochondrial différent et donc un taux d'hétéroplasmie variable entre elles. Le taux d'hétéroplasmie constitue une composante importante du phénotype clinique des patients (Bai and Wong, 2005).

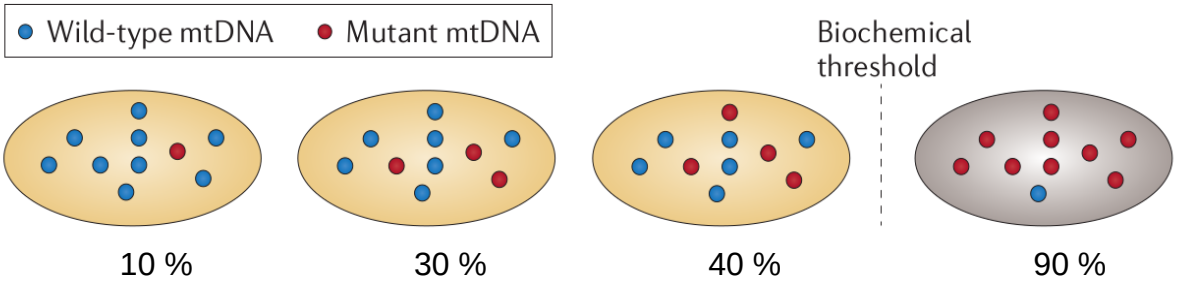


Figure 10 : Variation du taux d'hétéroplasmie
Le taux d'hétéroplasmie d'une cellule correspond au ratio du nombre de molécules d'ADNmt porteurs de la variation (rouge) divisé par le nombre total de molécule d'ADNmt (rouge + bleu). Les analyses biochimiques peuvent mettre en évidence des défauts dans la chaîne respiratoire à partir d'un seuil estimé aux environs de 80 % d'hétéroplasmie d'une variation pathogène (d'après (Stewart and Chinnery, 2015)).

L'ADN mitochondrial possède également un degré de polymorphisme important lié à un taux de mutation qui est jusqu'à 10 fois plus élevé que celui de l'ADN nucléaire. Ce polymorphisme est également à l'origine des haplogroupes mitochondriaux. Ces derniers sont définis par une combinaison de variations (par rapport à la séquence de référence mitochondriale rCRS) qui dépend de l'origine populationnelle des individus concernés. Il existe ainsi plus de 5000 combinaisons qui découlent de la mitochondrie « Eve », ancêtre commune des mitochondries transmises par lignée matrilinéaire (**Fig. 11**). La détermination de l'haplogroupe d'un patient est nécessaire. Les variations qui définissent un haplogroupe dépendent de l'origine ethnique du patient et peuvent ainsi être filtrées lors de l'analyse des variations présentes au sein de son ADNmt.

Enfin, au cours de l'évolution, une partie de l'ADN mitochondrial s'est intégrée au génome nucléaire provoquant l'apparition de régions pseudomitochondriales au sein de ce dernier. Ces régions ont une séquence très proche de celles de l'ADNmt et peuvent être capturées indirectement lors du séquençage d'exome.

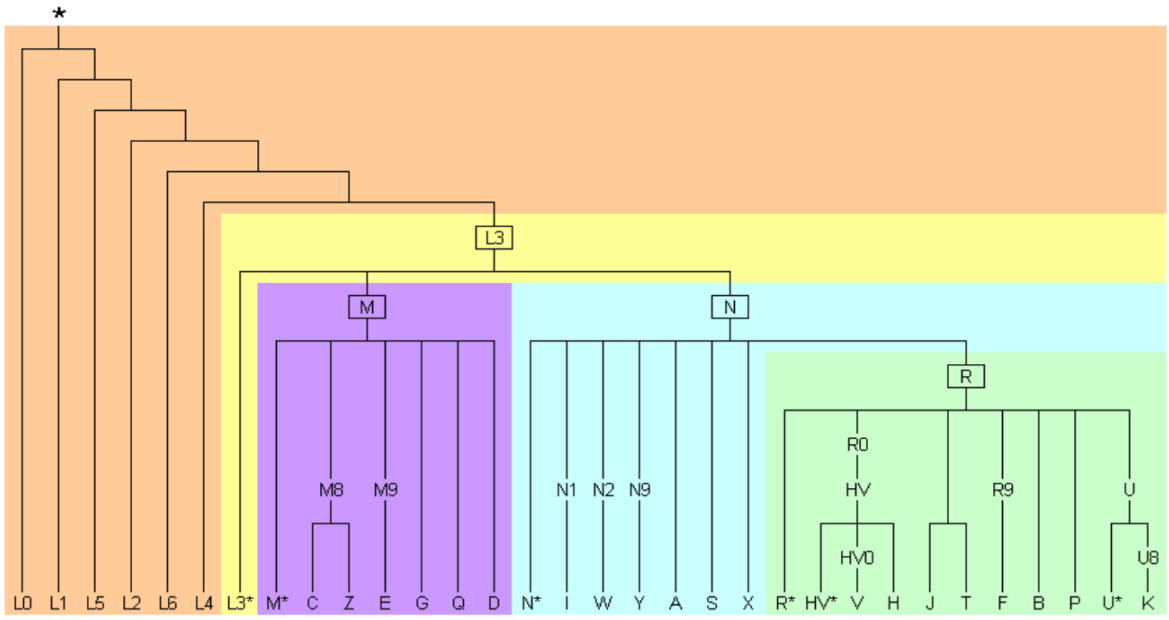


Figure 11: Représentation simplifiée de la « phylogénie » des haplogroupes mitochondriaux
L'ancêtre commun le plus récent par la lignée matrilinéaire (mitochondrie « Eve ») est figuré par une étoile. Les lettres suivies d'un astérisque représentent l'ensemble des haplogroupes qui en découlent (d'après (Van Oven and Kayser, 2009)).

I.3.2- Hérédité mitochondriale

Contrairement à l'ADN nucléaire qui a une origine biparentale, le génome mitochondrial a une hérédité exclusivement maternelle, même si de rares cas de transmission paternelle ont été décrits (Luo et al., 2018; Rius et al., 2019). Les mitochondries paternelles sont détruites dans le cytoplasme du zygote après fécondation (Sato and Sato, 2013). Ne subsistent alors que les mitochondries de la mère. Ce type de transmission a été confirmée en 1980 (Giles et al., 1980).

I.3.3- Les pathologies mitochondriales

Qu'elles soient d'origine nucléaire ou mitochondriale, les pathologies mitochondriales entraînent un dysfonctionnement de la chaîne respiratoire mitochondriale. Le spectre clinique

des pathologies mitochondriales est très variable (**Tableau 1**), peut débuter en période anténatale comme à un âge adulte avancé, ne concerner qu'un organe comme la surdité sensorineurale (OMIM 580000) ou plusieurs organes comme les muscles et le cerveau dans le syndrome de MERFF (épilepsie myoclonique avec fibres rouges déchiquetées OMIM 545000). Cet organite cellulaire est présent dans toutes les cellules. Cela explique que différents organes puissent être atteints de dysfonctionnements mitochondriaux causés par des mutations génétiques. Une centaine de gènes nucléaires et mitochondriaux sont impliqués dans le fonctionnement de cet organite (Dinwiddie et al., 2013) et peuvent présenter une variation associée à une pathologie. L'hétérogénéité des maladies mitochondriales est aussi présente au niveau génétique. Différentes variations peuvent être à l'origine de la même pathologie (ex : surdité sensorineurale) ; et plusieurs phénotypes (ex : MELAS OMIM 540000 ou CPEO OMIM 530000) peuvent avoir la même origine génétique.

Abréviations	Syndromes
Ataxia neuropathy	Ataxia neuropathy Syndromes
DEAF	Maternally inherited DEAFness or aminoglycoside-induced DEAFness
KSS	Kearns Sayre Syndrome
Leigh	Leigh Syndrome
LHON	Leber Hereditary Optic Neuropathy
LIMM	Lethal Infantile Mitochondrial Myopathy
MELAS	Mitochondrial myopathy, Encephalopathy, Lactic Acidosis and Stroke like episodes
MERRF	Myoclonic Epilepsy with Ragged Red Fibers
MIDD	Maternal Inherited Diabetes and Deafness
MNGIE	Mitochondrial Neurogastrointestinal Encephalopathy
NARP	Neuropathy, Ataxia and Retinis Pigmentosa
Pearson	Pearson Syndrome
PEM	Progressive EncephaloMyopathy
PEO	Progressive External Ophtalmoplégia
SNHL	SensoriNeural Hearing Loss

Tableau 1 : Principaux syndromes dus à des mutations de l'ADNmt

Devant cette hétérogénéité des maladies mitochondriales, l'identification de variations impliquées dans le phénotype de ces patients nécessite une analyse complète du génome mitochondrial.

II- SÉQUENÇAGE DE PREMIÈRE GÉNÉRATION : SANGER ET MICROARRAY

L'identification des causes génétiques des maladies rares repose le plus souvent sur le séquençage de l'ADN des patients concernés, éventuellement couplé à une recherche de grands réarrangements par d'autres techniques de cytogénétique moléculaire. Les techniques de séquençage ont évolué, du séquençage Sanger aux séquenceurs de deuxième génération.

II.1- Le séquençage par méthode Sanger

Cette méthode de séquençage de première génération, consistant en la détermination de l'ordre des bases nucléotidiques A, T, G et C, a été mise au point par Frederick Sanger, prix Nobel de chimie, en 1977. Elle se base sur le processus biologique de la réplication de l'ADN dans la cellule qui fait intervenir une enzyme : l'ADN polymérase. La méthode Sanger consiste en l'interruption de la synthèse enzymatique d'un nouveau brin à partir d'un brin préexistant qui sert de matrice ou modèle. Le fragment d'ADN à séquencer est d'abord amplifié par PCR. Sont ensuite ajoutés les amorces, l'ADN polymérase et un mélange de dNTP et ddNTP. Ces derniers, présents en plus faible concentration, sont des nucléotides dont le groupement OH des riboses a été remplacé par un hydrogène. Cette modification entraîne l'arrêt de la synthèse enzymatique du brin d'ADN néo-formé. L'intégration de ces ddNTP étant aléatoire lors de l'étape d'élongation, la terminaison de la réplication s'effectue de manière statistique sur toutes les positions possibles. Les fragments d'ADN obtenus à l'issue de la réaction sont donc de tailles différentes.

Initialement, la réaction se déroulait dans 4 mélanges différents contenant chacun un ddNTP radioactif (^{32}P). La lecture visuelle de la séquence du fragment d'ADN était effectuée par migration des 4 milieux de réaction sur gel d'électrophorèse. La distance de migration déterminait la position au sein de la séquence et la colonne donnait la nature du nucléotide. Environ 1000 paires de bases pouvaient ainsi être déchiffrées par lecture en 6-8 heures. Cette méthode nécessitait une lecture par échantillon. En 1990, les ddNTP fluorescents ont remplacé les ddNTP radioactifs. La méthode ne nécessite alors qu'un seul milieu de réaction. Cette modification de réactifs permet également la lecture de la séquence par séquenceur à capillaires à partir de 1999. Les fragments d'ADN migrent sur un gel avant de passer devant un laser. La nature du nucléotide marqué est déterminée par la longueur d'onde qu'il renvoie suite à l'excitation de son marqueur fluorescent par le laser. La vitesse de migration du fragment d'ADN dépend de sa taille et permet de déterminer la position, dans le fragment à séquencer, du ddNTP lu (**Fig. 12**). Il devient alors possible de lire 300 kb par lecture en 3h. De plus, 384 échantillons peuvent être traités en parallèle. Il s'agit d'un des outils principaux utilisé par le consortium *Human Genome Project* pour séquencer le génome humain en 2001 (Collins et al., 2003).

Mais cette méthode présente une limite importante qui est son faible rendement. Les différentes étapes de préparation sont longues et coûteuses et la taille des fragments est limitée à environ 1 kb. Une des alternatives développées est la puce à ADN mise au point pour détecter les variations ponctuelles.

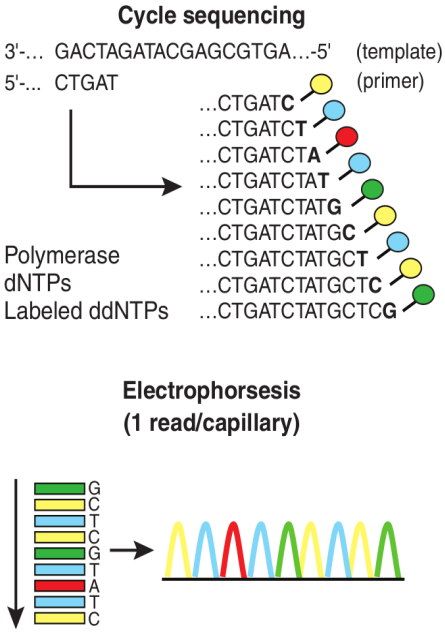


Figure 12 : Séquençage Sanger
Les copies du fragment d'ADN à séquencer sont de tailles différentes. La réaction de séquençage est stoppée par un ddNTP marqué d'un fluorochrome. Les fragments sont mis à migrer sur gel où leur vitesse dépend de leur taille. Le signal du premier fragment identifie le premier nucléotide de la séquence à lire (ici une cytosine). La séquence est décryptée à mesure que les fragments passent sous la caméra qui enregistre les signaux des ddNTP marqués et excités par un laser (pic de couleur) (d'après (Shendure and Ji, 2008)).

II.2- Puce à ADN (« SNP array ») et CGH-array

Initialement, cette méthode a été mise au point au début des années 90 afin de mesurer le niveau de transcription de milliers de gènes simultanément (puce à expression). Cette technique fait appel au principe d'hybridation par complémentarité de deux brins d'ADN ou d'un brin d'ADN sonde avec un ADN complémentaire issu d'un ARN. L'usage de cette méthode s'est ensuite élargi à la détection des Single-Nucleotide Polymorphisms (SNP) (**Fig. 13**). La puce à SNP est tapissée d'oligonucléotides simples brins, appelées sondes, de séquences connues : les différents allèles d'un SNP peuvent ainsi être représentés. L'ADN de l'individu à tester est d'abord fragmenté puis marqué par un fluorochrome avant d'être mis à incuber sur la puce. Il y a alors hybridation entre les fragments d'ADN marqués et leur brin complémentaire fixé à la puce.

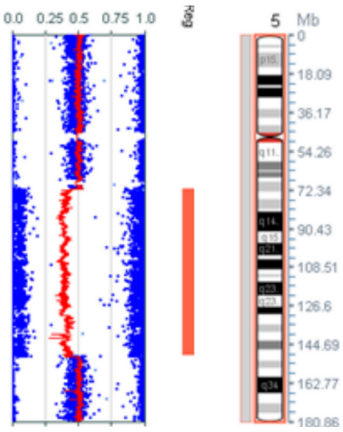


Figure 13 : Détection d'une délétion 5q13.2-5q33.1 par SNP array
Les génotypes sont représentés en bleu avec les homozygotes à gauche et à droite et les hétérozygotes au centre. La perte d'hétérozygotie indique une délétion hétérozygote de la région concernée (d'après (MacKinnon et al., 2012)).

Il existe deux limites principales à cette technique : l'une technique avec l'hybridation non spécifique et l'autre dans le champ de l'existence de variations non connues (i) dans un des gènes ciblés par la puce ou (ii) dans une région non étudiée. En cas de forte homologie entre des séquences d'ADN, une sonde risque de s'hybrider de façon non spécifique avec l'ensemble de ces fragments. Il peut donc y avoir un biais dans le signal détecté. Par ailleurs, la puce à ADN ne peut détecter les variations ponctuelles qu'au niveau des gènes prévus dans l'analyse (Bumgarner, 2013). Ainsi, si un individu présente un SNP supplémentaire non connu ou une variation causale dans une région non ciblée, le gène responsable de la pathologie ne sera pas identifié.

Les puces à SNP sont une variante de la CGH (« Comparative Genomic Hybridization »). Cette dernière est une méthode faisait appel à l'hybridation génomique pour détecter les anomalies chromosomiques déséquilibrées que sont les CNV. Des sondes de 60-80 pb sont fixées sur une puce ou « CGH-array ». L'ADN du patient et un ADN contrôle sont marqués par deux fluorochromes de couleurs différentes. Ils sont ensuite déposés en quantité identique et simultanément sur la puce. Les fluorochromes sont excités par leur longueur d'onde respective et l'intensité de la fluorescence est mesurée. Pour chaque sonde, le ratio entre les 2 intensités donne le rapport entre le nombre de copies d'ADN témoin et le nombre de copies d'ADN du patient. Chaque valeur de ratio peut être représentée sur un idéogramme de

chaque chromosome qui permet la visualisation des CNV (**Fig. 14**).

Contrairement aux puces à SNP, seules les anomalies déséquilibrées peuvent être détectées. Les pertes d'hétérozygotie ou l'origine parentale des remaniements ne peuvent pas être identifiées.

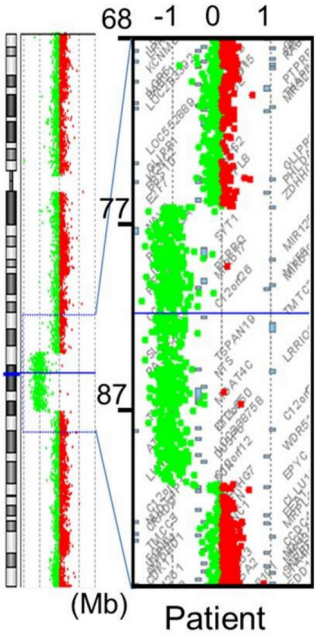


Figure 14 : Détection d'une délétion 12q21.2-q21.33 de 14 Mb par CGH-array
L'ADN témoin a été marqué en vert et l'ADN du patient en rouge. La région 12q21.2-q21.33 ne présente qu'une fluorescence verte témoin de la délétion de ce fragment chez le patient (d'après (Matsumoto et al., 2014)).

Les techniques de séquençage de première génération permettent une analyse ciblée de régions génomiques ou de gènes connus. Ces méthodes deviennent insuffisantes pour une étude d'exome ou de génome complet. Il a fallu attendre les séquenceurs de deuxième génération pour repousser ces limites, pour permettre de séquencer l'ensemble de nos gènes, et augmenter la probabilité d'identifier la ou les variations causales pour un individu donné.

II.3- Particularités de l’étude moléculaire de l’ADNmt

Avant l’avènement du séquençage à haut débit, l’étude moléculaire, c’est-à-dire l’identification des grands réarrangements et des variations ponctuelles de l’ADNmt, était réalisée par des méthodes de biologie moléculaire ciblant spécifiquement l’ADNmt, ses gènes et/ou ses variations en se basant sur le phénotype du patient. De nos jours, les grands réarrangements (grandes délétions ou grandes duplications) sont toujours mis en évidence par Southern Blot ou par PCR longues dont les produits sont mis à migrer sur gel d’agarose révélé aux UV (**Tableau 2**). La déplétion de l’ADNmt (diminution du nombre de molécules d’ADNmt) est quant à elle identifiée par la technique de PCR quantitative en temps réel. Les variations ponctuelles les plus récurrentes sont mises en évidence par PCR-RFLP (Holt et al., 1990). Les variations ponctuelles moins fréquentes sont identifiées par séquençage Sanger. Si la première méthode permet de déterminer précisément le taux d’hétéroplasmie, la deuxième ne donne qu’une estimation.

Lorsque l’analyse de l’ADNmt par ces méthodes ne permet pas l’identification de la variation causale, le reste de la séquence mitochondriale peut être étudiée. Le séquençage des 16 kb de cet ADNmt est réalisable par Sanger mais la méthode est fastidieuse. Des techniques moléculaires plus adaptées à cette étude ont donc été mises au point : la méthode de reséquençage Affymetrix (Maitra et al., 2004) qui permet d’identifier les variations homoplasmiques, et la méthode Surveyor® (Bannwarth et al., 2005) qui est efficace pour déterminer le taux d’hétéroplasmie. Ces méthodes d’analyse moléculaire de l’ADNmt sont longues et coûteuses quand il s’agit d’analyser le génome mitochondrial dans son intégralité. Le séquençage à haut débit est une technique qui permet de palier cette limite. Deux méthodes existent.

La première méthode est dite directe, c’est-à-dire que le génome mitochondrial est la « cible » de la méthode utilisée. L’ADNmt est enrichi à partir de l’ADN cellulaire. Pour cela il est possible d’utiliser la technique de « microarray » (Vasta et al., 2009) pour capturer l’ADN mitochondrial. Cet ADN est ensuite amplifié par PCR avec plusieurs couples d’amorces. D’autres méthodes, plus récentes, proposent une amplification PCR ne faisant appel qu’à un seul couple d’amorces (Zhang et al., 2012; Cui et al., 2013). Cela permet d’obtenir une

couverture plus uniforme et moins d'interférences avec les régions pseudomitochondriales présentes au sein du génome nucléaire. L'ADN est ensuite séquencé par la méthode d'Illumina.

Le séquençage indirect du génome mitochondrial consiste à obtenir les données à partir de celles déjà existantes. En effet, les régions pseudomitochondriales ciblées par les kits de capture ont des séquences très proches de celles de l'ADNmt ce qui conduit à la capture indirecte de ces dernières (**Fig. 15**). Il a été démontré que l'intégralité de la séquence de l'ADNmt peut être reconstituée à partir de ces données indirectes (Picardi and Pesole, 2012). La couverture du génome mitochondrial est alors dépendante de sa proportion dans l'extrait d'ADN cellulaire.

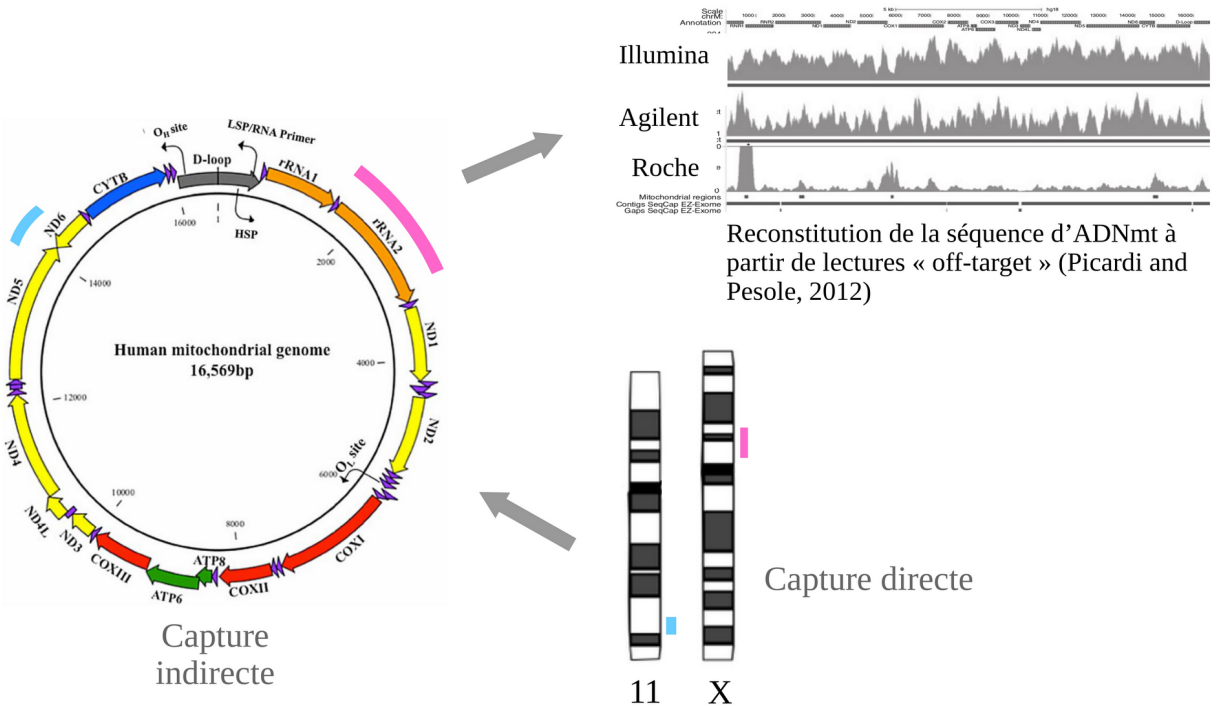


Figure 15 : Capture indirecte de l'ADNmt lors de l'ES
La séquence des régions exoniques du génome nucléaire sont obtenues par capture directe. Les sondes des régions pseudomitochondriales peuvent s'hybrider avec des régions de l'ADNmt conduisant au séquençage de ces dernières. L'ensemble des lectures « off-target » permet de reconstituer la séquence de l'ADNmt. La couverture dépend du kit de capture.

À l'instar des variations nucléaires identifiées par l'analyse d'exome, les variations mitochondriales suspectées d'être causales doivent être validées par une seconde technique moléculaire. Dans le cas de l'ADNmt, les méthodes employées sont le séquençage Sanger, la PCR-Surveyor® ou la PCR-RFLP ; qui déterminent également l'état homo- ou hétéroplasmique voire le taux d'hétéroplasmie lorsque leur sensibilité le permet.

Méthodes	Éléments détectés	Références
PCR longues	Grandes délétions et duplications	(Hu et al., 2007)
Southern Blot	Grandes délétions et duplications	(Southern, 1975)
Sanger	Variations	(Sanger and Coulson, 1975)
PCR-RFLP	Variations	(Holt et al., 1990)
Reséquençage Affymetrix	Variations homoplasmiques	(Maitra et al., 2004)
Surveyor ®	Variations hétéroplasmiques	(Bannwarth et al., 2005)
PCR quantitative en temps réel	Variations hétéroplasmiques et déplétions	(Bai and Wong, 2004)
Séquençage haut débit direct	Génome mitochondrial entier	(Vasta et al., 2009) (Cui et al., 2013) (Zhang et al., 2012)
Séquençage haut débit indirect	Génome mitochondrial entier	(Picardi and Pesole, 2012)

Tableau 2 : Méthodes de détection de l'ADN mitochondrial

III- SÉQUENÇAGE DE DEUXIÈME GÉNÉRATION : L'AVÈNEMENT DU SÉQUENÇAGE À HAUT DÉBIT D'EXOME ET DE GÉNOME

En 1989 débute le projet génome humain (« Human Genome Project »), destiné à séquencer l'intégralité du génome humain. Le consortium public international qui a collaboré sur ce programme a publié une première ébauche de la séquence du génome en février 2001 (International Human Genome Sequencing Consortium, 2001). La séquence complète fut disponible en avril 2003. Le consortium public international a choisi de séquencer le génome humain par la méthode de shotgun hiérarchique. Cette méthode, imaginée par Frederick Sanger, permet de séquencer des génomes entiers. Elle consiste en la fragmentation aléatoire du génome. Les fragments (grands clones) de 100-200 kb sont insérés dans des vecteurs (chromosomes artificiels de bactérie ou BAC) eux-mêmes intégrés dans des bactéries *E.coli*. Des marqueurs présents au sein de ces grands clones permettent de les ordonner le long des

chromosomes humains et d'établir ainsi la carte physique du génome humain. Chaque grand clone est ensuite fragmenté en segments d'environ 1000 pb qui sont ensuite séquencés par la méthode Sanger. La recherche de séquences communes entre ces petits clones permet de reconstituer la séquence complète de chaque grand clone. L'ensemble des séquences des grands clones localisées sur la carte physique fournit la séquence du génome humain.

L'obtention de la séquence complète du génome humain a constitué une avancée majeure pour la recherche, notamment en biologie moléculaire. Cette même décennie a aussi connu l'essor du séquençage de deuxième génération permettant un séquençage massif, rapide et moins coûteux de génomes entiers. La recherche des variations impliquées dans les maladies génétiques s'est concentrée dès le début sur les régions codantes du génome humain, également appelées exome, qui correspond à environ 2 % du génome. Le séquençage à haut débit d'exome génère une masse de données à analyser et à stocker moins importante que pour un génome complet et à moindre coût. De plus, on estime que 85 % des variations pathogènes causales sont exoniques (Majewski et al., 2011). Le séquençage d'exome se déroule comme celui du génome à l'exception d'une étape de capture des régions d'intérêt (exons) en amont de la réaction de séquençage. Les résultats du premier séquençage d'exome publiés en 2009 a permis l'identification d'un nouveau gène impliqué en pathologie humaine (Ng et al., 2009). Au cours de la décennie suivante, les technologies de séquençage ont continué de s'améliorer, diminuant toujours plus les coûts de séquençage.

III.1- Le séquençage à haut débit d'exome

II.1.1- Les technologies de séquençage à haut débit d'exome

Comme pour la méthode Sanger, le séquençage d'exome en short-reads (ou fragments courts) utilise également le principe de réplication de l'ADN.

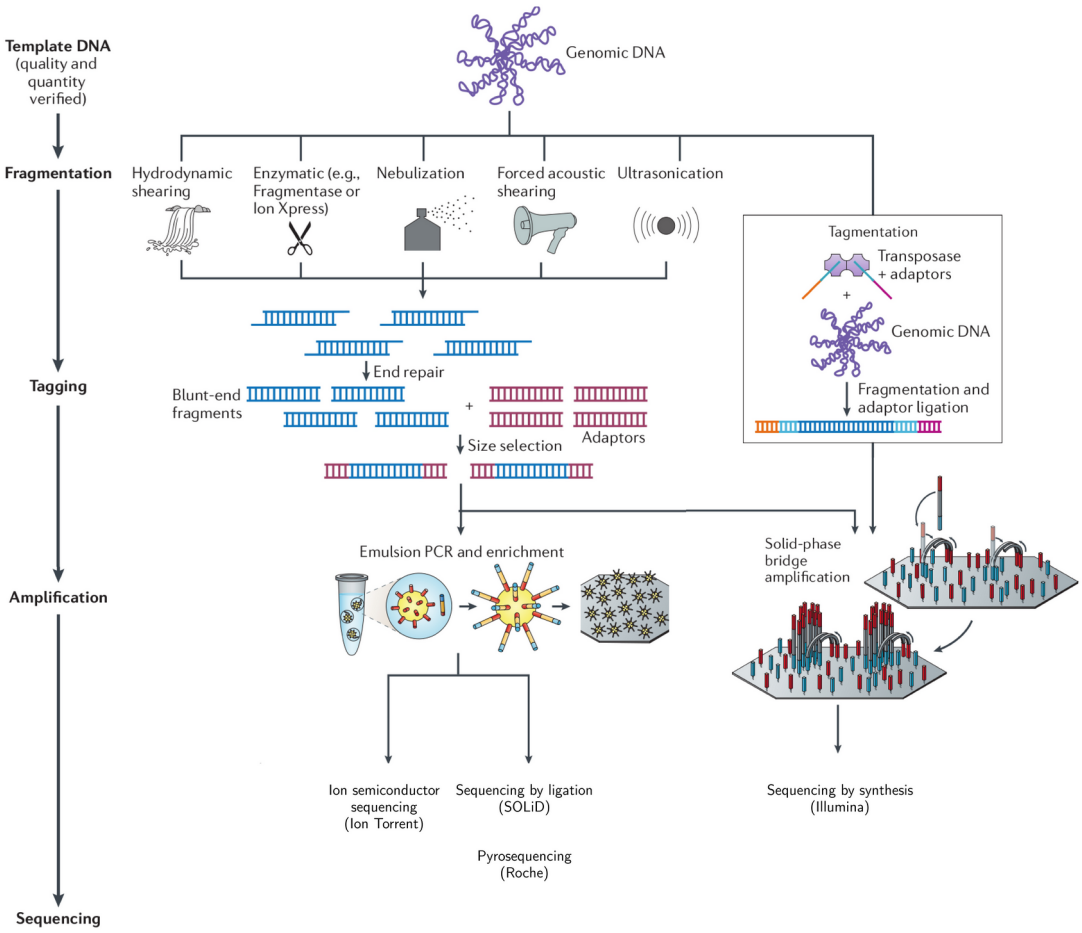


Figure 16 : Les différents protocoles de séquençage à haut débit d'exome
Seule la technologie Illumina utilise l'amplification par « bridge-PCR ». Les technologies de SOLiD et Roche font appel à l'amplification par PCR en émulsion et à une détection optique des signaux de séquençage. La technologie de Ion Torrent diffère des deux précédentes en détectant des variations de pH (d'après (Loman et al., 2012)).

La première étape du séquençage à haut débit d'exome est la fragmentation de l'ADN d'intérêt par action physique, chimique ou enzymatique (**Fig. 16**). Des adaptateurs sont ensuite fixés aux deux extrémités de chaque fragment par une ADN ligase. Ces petites séquences nucléotidiques contiennent les amorces de séquençage et la séquence permettant la fixation sur le support d'amplification ou de séquençage. Dans le cas du séquençage spécifique des régions codantes (ou exome), le processus nécessite une étape supplémentaire de capture des régions exoniques, appelée enrichissement. Les exons capturés sont dépendants du kit de capture utilisé. En ce qui concerne la technologie Agilent, une des plus utilisées, cette étape utilise des fragments d'ARN biotinylés contenant les séquences d'intérêt prédéfinies appelées sondes. Les sondes à ARN ont été choisi plutôt que les sondes à ADN car la force de liaison d'un duplex

ARN-ADN (liaison des nucléotides A::U) est plus importante que celle d'un duplex ADN-ADN (liaison A::T). Le principe de la capture consiste en l'hybridation de ces ARN avec les fragments d'ADN que l'on souhaite séquencer. Les duplex ainsi formés sont retenus par des billes recouvertes de streptavidine qui lie la biotine des sondes. Après élimination des fragments non retenus, l'ARN est digéré. Les fragments d'ADN, alors simples brins, sont ensuite amplifiés. Cette amplification peut être réalisée avec plusieurs techniques dont notamment la PCR sur support plan (« bridge PCR », **Fig. 17a**) et la PCR en émulsion (**Fig. 17b**). L'ADN d'intérêt peut ensuite être séquencé.

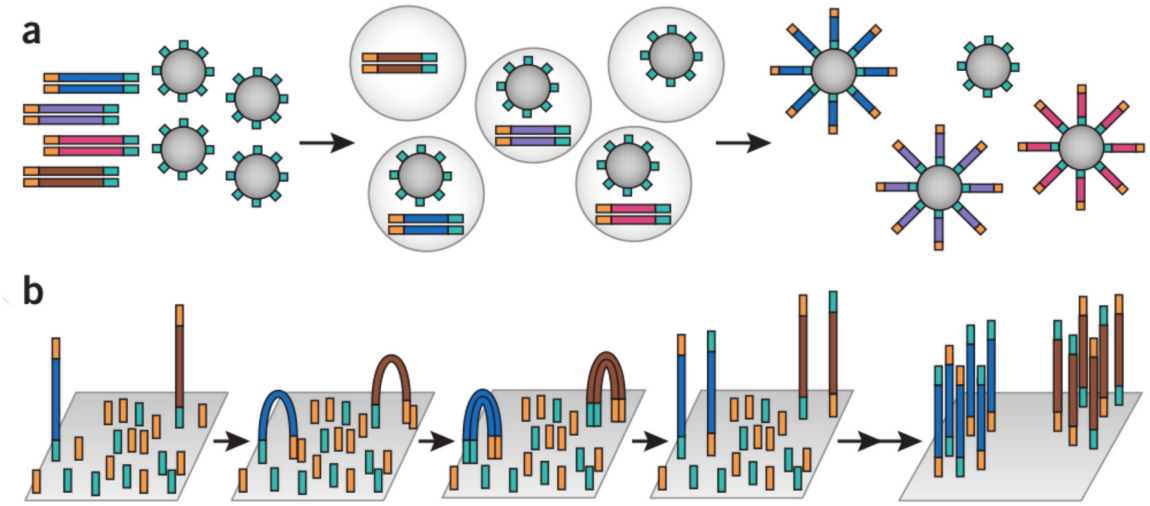


Figure 17 : Amplification par PCR des fragments d'ADN à séquencer
a) PCR en émulsion. L'ADN est fractionné aléatoirement et des adaptateurs spécifiques des extrémités 5' et 3' sont ajoutés. L'ADN est ensuite fortement dilué et placé en contact des billes. La dilution importante de l'échantillon permet la liaison d'un seul fragment d'ADN par bille, grâce à la séquence complémentaire de l'adaptateur présente sur la bille. Chaque bille va ainsi permettre d'amplifier un seul fragment d'ADN. Elles vont ensuite être émulsionnées avec un mélange d'eau et d'huile contenant les produits d'amplification PCR : ce sont des microréacteurs. Les fragments d'ADN sont amplifiés individuellement dans leur microréacteur. A la fin du processus, les billes sont recouvertes de copies du même fragment. b) « Bridge-PCR ». L'ADN est fractionné en fragments doubles-brins auxquels sont ajoutés des adaptateurs spécifiques des extrémités. Après dénaturation, les fragments simples brins se lient par une extrémité à la flow-cell recouverte d'adaptateurs. Le changement des conditions du milieu permet la fixation de la seconde extrémité. Les fragments forment alors des ponts qui permettent la réplication à chaque cycle. Les fragments sont ainsi amplifiés par groupes ou clusters (d'après (Leong et al., 2014)).

Il existe quatre techniques principales de séquençage à haut débit : le pyroséquençage (454 de Roche), le séquençage par synthèse (Illumina), le séquençage par ligation (SOLiD) et le séquençage par semi-conducteur (Ion Torrent).

Le pyroséquençage a été développé par Pål Nyrén et ses collègues de l'Institut Royal de Technologie de Stockholm au début des années 90 (Nyren et al., 1993) et est basé sur la détection de pyrophosphate. Commercialisée en 2005 par la société 454 Life Science (rachetée ensuite par Roche), il s'agit de la première méthode utilisée en séquençage haut débit. Elle a permis le séquençage, par la technologie haut débit, du premier génome humain complet en 2008. Les fragments simples brins d'intérêt ont été amplifiés par PCR en émulsion. Chaque bille, qui porte une séquence unique amplifiée, est ensuite déposée dans un puits avant l'ajout des enzymes de séquençage et des amorces. Lors de la synthèse du brin complémentaire, l'incorporation d'un nouveau nucléotide est suivie de la libération d'une molécule de pyrophosphate inorganique. Ce dernier est ensuite convertit en ATP par l'ATP-sulfurylase. La luciférase va alors le coupler à une luciférine. Cette réaction produit de l'oxyluciférine et un signal lumineux capté par une caméra Charge-Coupled Device (CCD) qui le transforme en pic visible sur un pyrogramme. Plus il y a dNTP incorporés, plus le signal est lumineux, et donc plus la hauteur du pic est importante. Les dNTP sont introduits les uns après les autres après lavage. La plaque de séquençage étant fixe, il est possible de superposer les clichés et d'identifier chacun des fragments tout au long de la réaction de séquençage. En comparant l'ordre des dNTP introduits et la présence ainsi que l'intensité de la réaction lumineuse il est possible de reconstituer la séquence d'ADN du fragment en cours de synthèse (**Fig. 18**).

En 1998 Ronaghi et son équipe améliorent la technique en éliminant les étapes de lavage intermédiaires (Ronaghi et al., 1998). En effet, l'ajout d'une enzyme comme l'apyrase, dégradant les nucléotides et les molécules d'ATP, permet de conserver le milieu de réaction tout au long du séquençage. Puis en 2005, Margulies et ses collaborateurs apportent une nouvelle amélioration en développant la méthode sur support solide de type micropuce (Margulies et al., 2005).

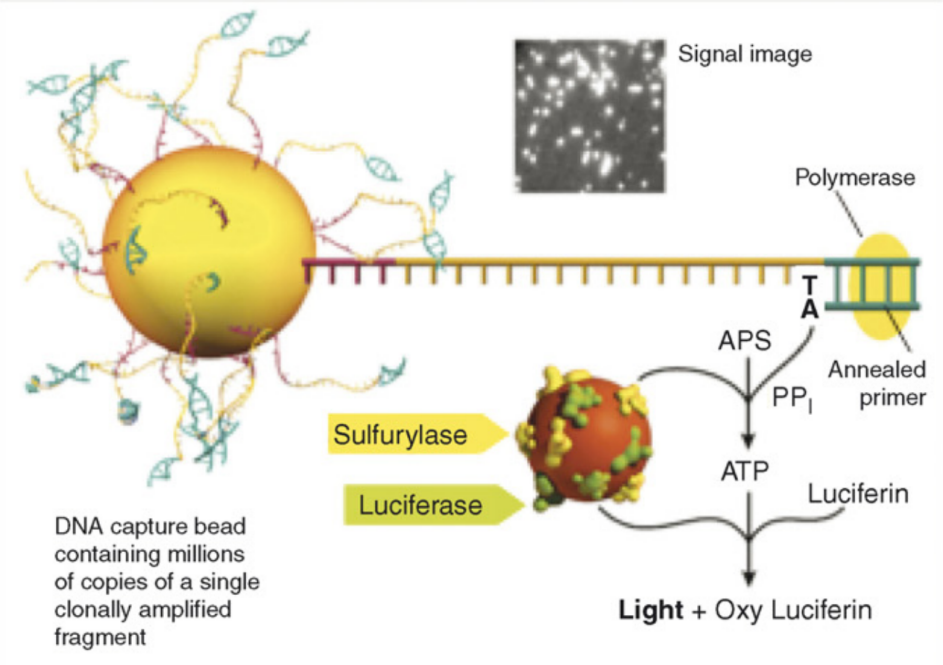


Figure 18 : Pyroséquençage
Chaque bille permet le séquençage d'un fragment unique. La libération simultanée de pyrophosphates inorganiques permet d'atteindre le seuil de détection du signal lumineux. L'ensemble des billes est séquencé simultanément. Chaque cliché détecte donc l'ensemble des signaux lumineux de la puce (d'après (Mardis, 2008)).

Cette technique s'avère rapide (un run en 7h). La longueur des fragments séquencés est de 400 pb. Les principaux inconvénients de cette méthode sont le prix élevé des réactifs et un taux d'erreur de 1 %. Les erreurs majeures lors du séquençage proviennent des homopolymères de taille supérieure à 6. Depuis 2016, cette technologie n'est plus maintenue.

Le séquençage par synthèse à l'aide de terminateurs réversibles a été développé au début des années 2000 par Shankar Balasubramanian et David Klenerman, fondateurs de Solexa (rachetée depuis par Illumina). Publiée en 2008, cette technologie permet de séquencer 1 Gb par run (Bentley et al., 2008) et est la plus utilisée aujourd'hui. L'ADN d'intérêt a été fragmenté en amont en morceaux allant jusqu'à 250 pb et amplifié par bridge PCR sur flow-cell. La réaction de séquençage se déroule directement sur ce support d'amplification ainsi recouvert de clusters monoclonaux de fragments d'ADN simple brin. La réaction consiste en l'utilisation de terminateurs réversibles qui sont des dNTP fluorescents. Leur extrémité 3' OH est bloquée par la liaison à un fluorophore, spécifique de chaque nucléotide. L'incorporation du

dNTP suivant est donc inhibée tant qu'il n'y pas clivage enzymatique du fluorophore. A chaque nouveau cycle, l'ADN polymérase et les terminateurs réversibles sont ajoutés sur la flow-cell. Un seul dNTP est alors incorporé à chaque brin néo-synthétisé et la réaction de séquençage est bloquée simultanément pour chaque fragment de la flow-cell. A chaque cycle, les fluorophores sont excités successivement par chacune de leurs longueurs d'onde. La fluorescence émise après le passage du laser est capturée par une photographie. La première base de chaque cluster peut ainsi être déterminée en fonction de la longueur d'onde émise. Les fluorophores sont ensuite clivés, la flow-cell est lavée et un nouveau cycle peut débuter. Les cycles s'enchaînent de façon à lire simultanément la succession des bases de chaque cluster de la flow-cell (**Fig. 19**). Cette méthode permet de lire les homopolymères et présente un taux d'erreur d'environ 0,1 %. Néanmoins le taux d'erreur de séquençage augmente avec la longueur des fragments. En effet, à mesure que la réaction de séquençage progresse, les lavages sont moins performants entraînant une augmentation du bruit de fond. De même l'ADN, polymérase perd en efficacité et les fluorophores leur fluorescence. Les erreurs dues au « pre-phasing » et au « phasing » s'accumulent également avec la longueur du fragment réduisant la qualité du séquençage. Le « pre-phasing » a lieu lorsqu'un fluorophore n'est pas éliminé par l'étape de lavage et qu'il sera de nouveau lu au cycle suivant. Le séquençage de ce fragment sera donc en retard d'un nucléotide par rapport au reste du cluster et son signal interférera avec ceux des fragments alentours. Le « phasing » se produit en cas d'anomalie d'un terminateur réversible qui conduit à l'incorporation de deux nucléotides au cours du même cycle. Il y aura également des interférences de signaux et le séquençage du fragment sera alors en avance d'une base par rapport au cluster.

En 2012, afin de diminuer la durée de séquençage, la compagnie Illumina choisit de réduire la durée de capture du signal en diminuant la surface à imager (Buermans and den Dunnen, 2014). Mais cette modification restreint la quantité de données générées par run augmentant ainsi le prix du séquençage par nucléotide. Ainsi, en 2014, Illumina introduit les flowcells « à motifs » qui contiennent des milliards de nanopuits. Cette structure permet d'accroître la densité des clusters de fragments et donc de diminuer le prix du séquençage.

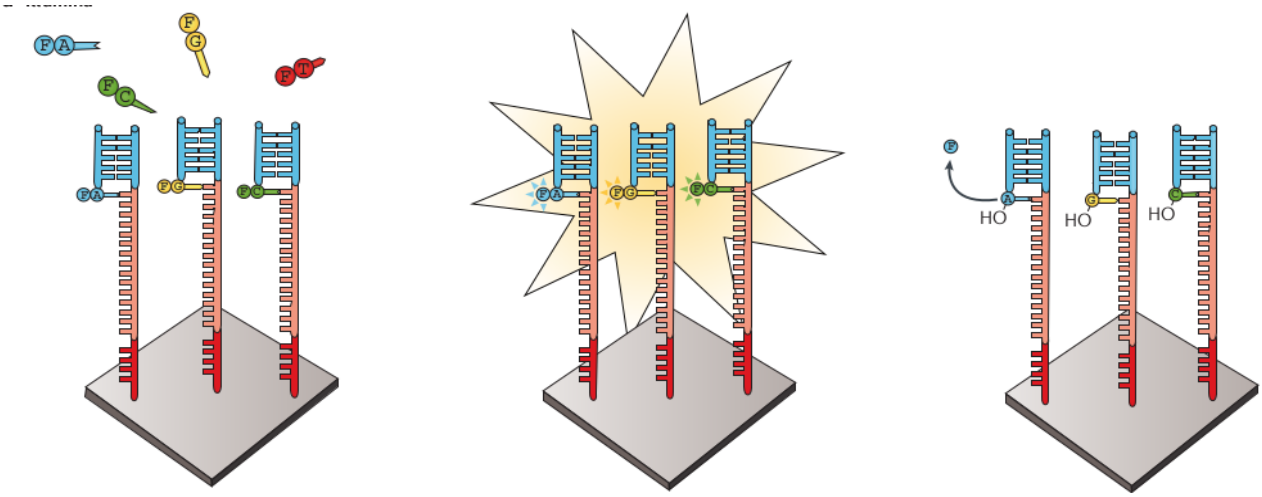


Figure 19 : Séquençage par terminaison réversible (technologie Illumina)
Les nucléotides marqués (ou terminateurs réversibles) sont ajoutés avec les produits de réaction. Le nucléotide complémentaire de la séquence s'hybride stoppant la réaction de séquençage (illustration de gauche). Les fluorophores sont excités par un laser et la fluorescence émise est enregistrée (illustration du milieu). Les fluorophores sont clivés et la flow-cell est lavée (illustration de droite). Un nouveau cycle peut débuter (d'après (Goodwin et al., 2016)).

Le séquençage par ligation, publiée en 2005 par George Church, est utilisé par la technologie SOLiD (Sequencing by Oligonucleotide Ligation and Detection) commercialisée par Life Technologies (acquise depuis par Thermo Fisher Scientific) en 2007. A l'instar du pyroséquençage, la réaction de séquençage de la technologie SOLiD se fait sur billes après une PCR en émulsion. Mais contrairement à la méthode 454, les billes sont fixées sur une plaque en verre, l'enzyme utilisée est une ligase au lieu de l'ADN polymérase et le signal détecté n'est pas une bioluminescence mais une fluorescence. Le séquençage par ligation fait appel à des sondes de 8 bases : les 2 premières bases sont connues, les 3 suivantes sont universelles afin de s'hybrider à n'importe quelles séquences et les 3 dernières, également universelles, sont couplées à un fluorochrome. Les 16 sondes (2^4 combinaisons) sont donc marquées par 4 fluorochromes. Ces derniers vont ainsi identifier 4 sondes chacun. L'information de séquençage obtenue sera donc partielle. C'est la répétition des cycles, en décalant les amorces d'une base à chaque fois, qui va permettre de préciser la séquence, et de lire 2 fois chaque base. Le séquençage d'un fragment complet se déroule en 5 cycles (**Fig. 20**).

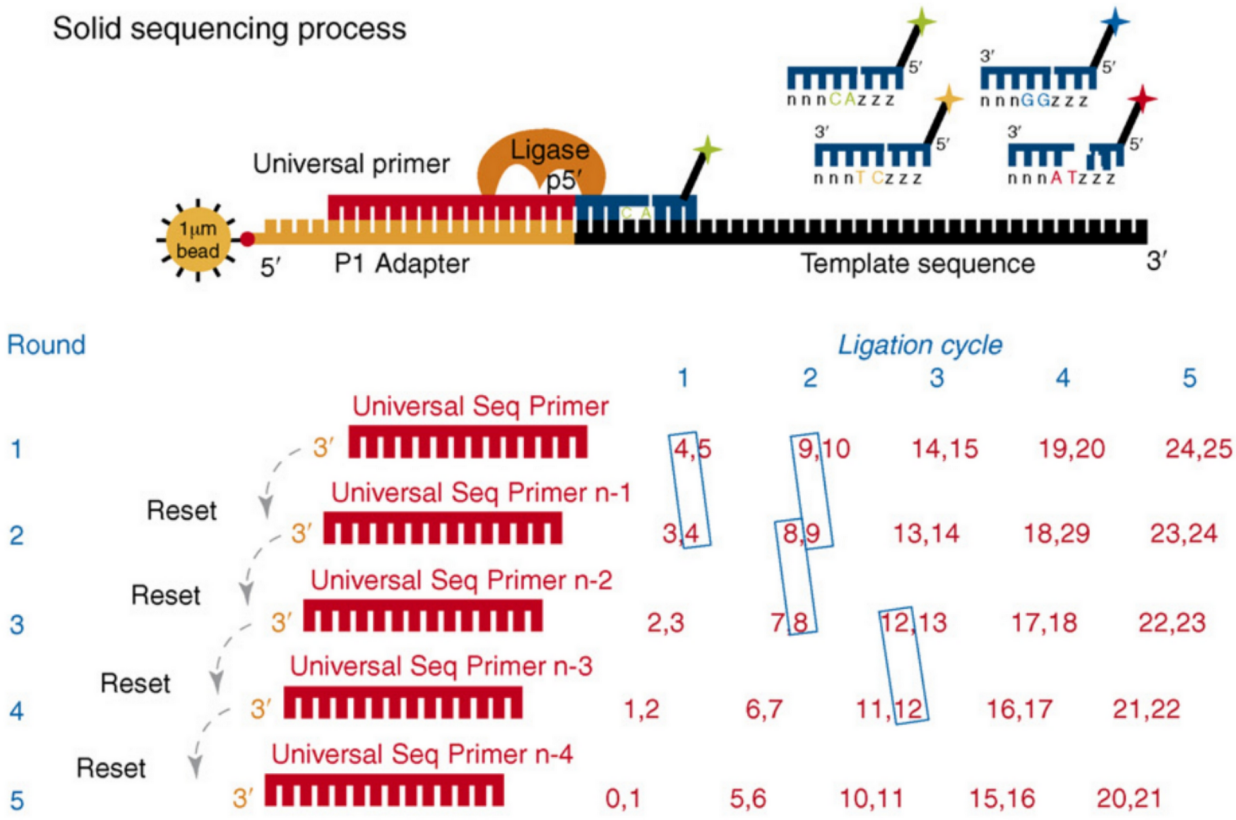


Figure 20 : Séquençage par ligation (technologie SOLiD)
Chaque sonde est composée de 8 nucléotides : 2 bases connues, 3 bases universelles et 3 autres bases universelles liées à un fluorochrome. Ces dernières sont clivées après ligation et enregistrement du signal fluorescent. Les bases sont identifiées par paires toutes les 5 positions. 5 cycles complets sont nécessaires pour reconstituer la séquence nucléotidique (d'après (Mardis, 2008)).

Lors du premier cycle, les amorces s'hybrident à une position n des adaptateurs. Les 16 octamères sont introduits et l'un d'entre eux va s'hybrider sur le brin matrice par complémentarité des deux premières bases et par intervention de la ligase. Les trois dernières bases de l'octamère empêchent la fixation d'une nouvelle sonde. Le fluorochrome est alors excité par un laser et le signal enregistré. Les trois dernières bases de la sonde qui portent le fluorochrome sont ensuite clivées et une nouvelle sonde peut venir s'hybrider 5 positions en aval de la précédente. Ce processus est répété de façon à séquencer l'intégralité du brin, pour chaque bille fixée sur la lame. Le brin néo-synthétisé est ensuite éliminé et une nouvelle amorce vient s'hybrider à la position n-1. Un nouveau cycle commence. La combinaison des codes de couleur à l'issue des 5 cycles permet d'obtenir la séquence complète du fragment d'ADN. Le double séquençage a conduit à la diminution du taux d'erreur (1/1000) et permet le traitement

de 256 échantillons en parallèle. Néanmoins, cette méthode ne permet le séquençage que de fragments de taille inférieure à 100 pb (Liu et al., 2012) et rencontre des difficultés pour les séquences palindromiques. En 2013, l'amplification sur billes a été remplacée par une amplification directe sur la puce, ce qui a permis d'augmenter le rendement de cette technologie (Ma et al., 2013). L'analyse bioinformatique des données séquencées par cette méthode était complexe. Depuis fin 2017, cette technologie n'est plus maintenue.

Le séquençage par semi-conducteur, commercialisé en 2010 par Ion Torrent systems, ne fait pas appel à la détection d'un signal lumineux mais à la mesure de variations locales du pH (Rothberg et al., 2011). Il s'agit de la première technologie à ne pas faire appel aux nucléotides marqués et à la détection optique ce qui permet de s'affranchir de la caméra. Dans les technologies de séquençage, l'imagerie représente l'étape limitante. En éliminant cette dernière, la méthode de séquençage par semi-conducteur augmente sa vitesse. Cette technologie utilise une puce semi-conductrice présentant des puits à sa surface. Les fragments à séquencer sont initialement amplifiés par PCR en émulsion billes. Chaque bille est ensuite insérée de façon unique dans un puits contenant les amorces de séquençage et l'ADN polymérase. Les nucléotides utilisés pour la réaction de séquençage ne sont pas modifiés ni marqués. C'est la libération d'un proton H+ au cours de l'incorporation d'un dNTP qui est détecté par changement du pH du puits. Les dNTP sont ajoutés successivement dans les puits après une étape de lavage. A chaque cycle, si le dNTP présent au sein du puits s'hybride, le proton H+ libéré fait varier le pH local mesuré au fond du puits par la technologie semi-conducteur. La modification du pH est directement proportionnelle au nombre de protons H+ libérés et donc de dNTP incorporés. Les variations de pH successives sont représentées sous forme d'ionogramme (**Fig. 21**). Le taux d'erreur de cette technologie est estimé à 1 %. La principale difficulté rencontrée lors du séquençage provient des homopolymères. La technologie a évolué rapidement depuis sa mise sur le marché avec l'augmentation de la surface de la puce, de la densité de puits et de la longueur des fragments d'ADN (Buermans and den Dunnen, 2014).

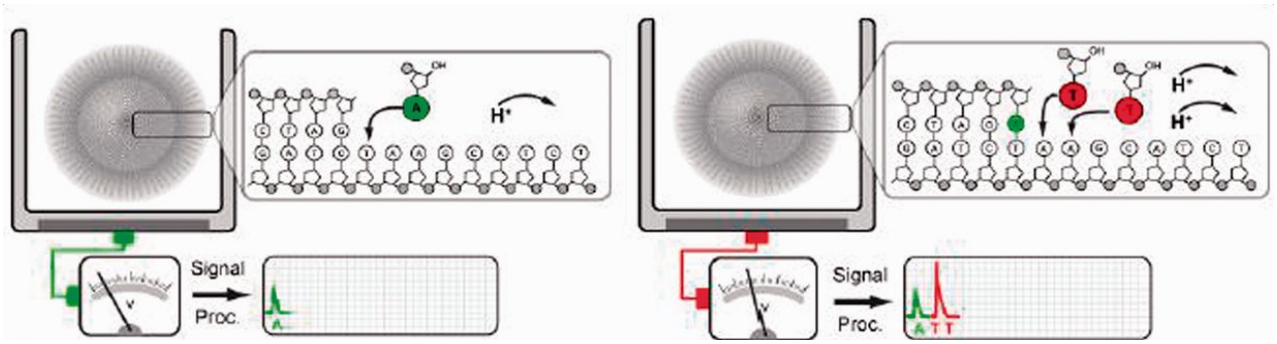


Figure 21 : Séquençage Ion Torrent
A chaque cycle, la libération d'un proton lors de l'incorporation d'un nucléotide fait varier le pH du puits. La technologie semi-conducteur au fond du puits détecte cette modification du milieu et le retranscrits sur un ionogramme. L'amplitude des pics dépend du nombre de nucléotides incorporés lors de la réaction de séquençage (Golan and Medvedev, 2013).

Les différents signaux détectés par ces technologies sont transformés en séquences nucléotidiques au format numérique. Ces données numériques nécessitent un traitement bioinformatique pour pouvoir être analysées.

II.1.2- L'analyse bioinformatique des données de séquençage à haut débit d'exome

La succession d'étapes de l'analyse bioinformatique est appelée « pipeline ». Le but est d'obtenir, à partir de données brutes de séquençage de l'échantillon, un fichier regroupant toutes les variations dans la séquence nucléotidique, dans le nombre de copies de segments chromosomiques ou dans la structure des chromosomes par rapport à une référence du génome humain. Il s'agit également d'annoter les variations et les gènes identifiés avec des informations issues de bases de données afin de faciliter la lecture des données de ES par le biologiste.

Les séquenceurs convertissent la séquence des fragments d'ADN en « reads » ou lectures (1 fragment = 1 read). Lorsque le séquençage n'a eu lieu que d'un seul sens de lecture on parle de reads classiques ou « single-end » (**Fig. 22**). À l'inverse, si les deux sens ont été lus, on parle de reads « paired-end ». Ce deuxième type de lecture améliore la précision notamment aux niveaux des régions répétées. L'ensemble des reads d'un échantillon est regroupé dans un fichier de données brutes FASTQ pour les reads classiques, ou dans deux fichiers (un par sens de lecture) pour le séquençage paired-end.

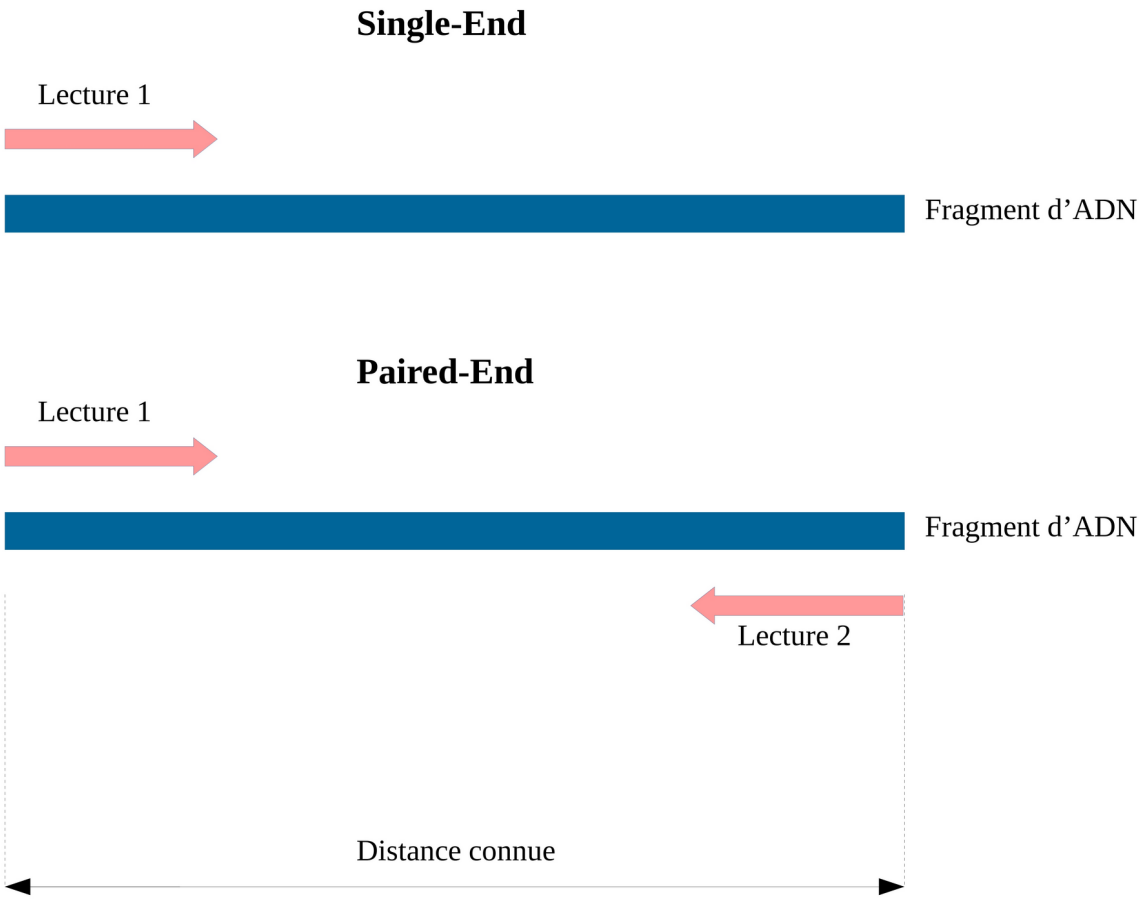


Figure 22 : Représentation schématique des séquençages en « single-end » et en « paired-end »
Lors du séquençage en « single-end » l'ADN n'est lu que dans un sens. Les lectures obtenues ont toutes la même orientation. Le séquençage en « paired-end » permet de lire l'ADN dans les deux sens. Chaque lecture orientée de 5' vers 3' (lecture 1) est appariée avec une lecture de sens opposée (lecture 2) dont la distance est connue.

Le format FASTQ contenant les données brutes de séquençage à haut débit est le format le plus répandu. Il est composé d'une succession de blocs de 4 lignes (un par lecture). La première renseigne sur l'identifiant unique du read concerné débutant systématiquement par le caractère « @ ». La deuxième est la séquence en elle-même. La troisième contient le symbole « + » et éventuellement l'identifiant du read. Enfin, la quatrième ligne contient les scores de qualité de chaque base, codés par un seul caractère ASCII. Ce score, appelé score Phred Q définit par $Q = -10 \log_{10} P$, où P est la probabilité que la base appelée soit une erreur. Afin de coder cette information sous la forme d'un seul caractère ASCII lisible par l'utilisateur, la valeur 33 est ajoutée au score. Avant Illumina 1.8 elle était de 64. La valeur ainsi obtenue est

reportée sur le code en base 10 de la table des caractères ASCII. Le caractère ainsi identifié est inscrit dans la quatrième ligne, sous la base concernée. Les fichiers FASTQ contiennent des données brutes. Pour augmenter la qualité des séquences il est nécessaire de les « nettoyer », c'est-à-dire d'éliminer les nucléotides de mauvaise qualité ou les séquences trop petites et de retirer les séquences correspondantes aux adaptateurs. Il s'agit de l'étape de « Trimming ».

La prochaine étape est l'alignement des lecrures sur le génome de référence afin de d'identifier la position des lectures sur le génome de référence. Ce processus est rendu possible par l'existence de plusieurs logiciels d'alignement, dont le plus utilisé pour son temps de calcul et sa précision est BWA (Li and Durbin, 2009). Le fichier d'alignement obtenu est un fichier SAM (pour Sequence Alignment/Map). Les positions génomiques de l'alignement, la séquence et le score Phred de chaque lecture alignée y sont sauvegardés. La taille de ce fichier pouvant devenir importante, il peut être compressé sous forme de fichier binaire de format BAM (pour Binary Alignment/Map). S'en suivent des étapes de marquage des duplicats PCR, de réalignement des indels et de recalibration des scores de qualité des bases afin d'améliorer la qualité des séquences.

Une fois le fichier BAM final obtenu, le « variant calling » ou appel de variant est réalisé afin d'identifier les variations de séquence du patient par rapport à la référence. Les résultats sont réunis dans un fichier VCF (pour Variant Caller Format). Il s'agit d'un format de fichier tabulé listant toutes les variations identifiées par rapport au génome de référence, ainsi que les informations bioinformatiques et biologiques de chacune de ces variations. Pour faciliter la lecture, les variations peuvent subir une ou plusieurs étapes d'annotations c'est-à-dire d'ajout d'informations spécifiques à chaque variation obtenue. L'annotation bioinformatique renseigne quant à l'expérience en précisant la profondeur, la couverture, le génotype, etc. L'annotation biologique donne des indications sur le nom des gènes impactés, les transcrits concernés, des informations de type base de données de fréquence populationnelle, score de prédiction de pathogénicité, annotations fonctionnelles, etc… Cette annotation permet de filtrer les variations. Par exemple, le rapport final qui est lu, peut ne contenir que les variations géniques non synonymes et de fréquence inférieure à 1 % dans la population générale (1000 Genomes Project Consortium et al., 2012).

II.1.3- Les limites du séquençage à haut débit d'exome

Le premier niveau de difficultés rencontrées dans l'étude de données de séquençage haut débit d'exome concerne **l'interprétation clinico-biologique** de cette analyse, sous-tendues avant tout par les avancées de la recherche produisant de nouvelles connaissances ou l'impact de l'hérédité des maladies rares (**Tableau 3**).

Les connaissances des maladies et des gènes au moment de l'analyse constituent une limite détectée dès les premières études de faisabilité du séquençage d'exome (Ng et al., 2010). En effet, de nombreuses variations sont mises en évidence au sein de gènes dont la fonction reste peu ou mal connue ce qui constitue un frein au diagnostic d'une maladie génétique. Néanmoins, les connaissances progressent grâce à la recherche avec en moyenne 20 gènes supplémentaires identifiés chaque mois ces 12 derniers mois, selon la base de données OMIM (OMIM - Online Mendelian Inheritance in Man., 1996) (**Fig. 23**). La réanalyse régulière permet ainsi d'augmenter d'environ 15 % le taux diagnostique (Nambot et al., 2018 ; Wright et al., 2018). Les systèmes de partage de données et de connaissances se sont avérés d'importance majeure, avec en particulier la base de données GeneMatcher (Bruel et al., 2019b).

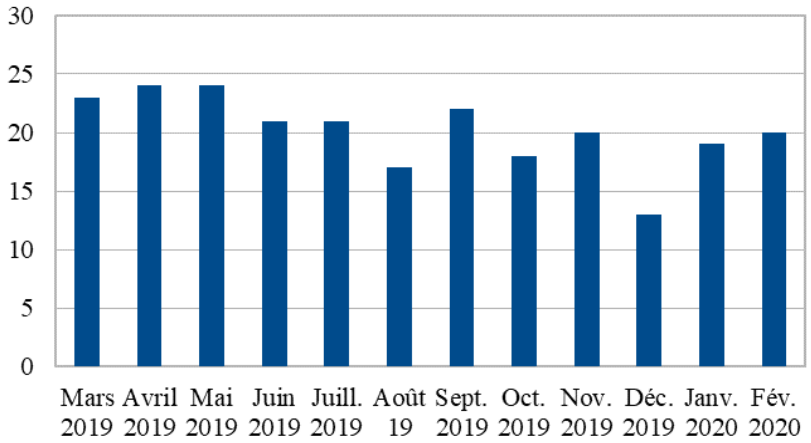


Figure 23 : Nombre de gènes nouvellement impliqués en pathologie humaine
242 gènes ont été identifiés ces 12 derniers mois comme impliqués dans des maladies génétiques. En moyenne, 20 gènes sont ainsi décrits chaque mois (adapté de (OMIM - Online Mendelian Inheritance in Man., 1996)).

Il existe également des cas "double-hit" où la présentation clinique du patient combine les symptômes de deux pathologies distinctes. Il est alors possible d'identifier les deux gènes responsables chacun d'une partie du phénotype.

Aux limites d'origine clinico-biologique, viennent s'ajouter des **difficultés d'analyse techniques**, qui proviennent de paramètres de la biologie moléculaire. On peut noter principalement la couverture incomplète de l'exome, souvent dans les régions riches en GC ou plus généralement les régions de faible complexité, ainsi que les pathologies en mosaïque (**Tableau 3**). En effet, les régions riches en GC ont une couverture plus faible que le reste de l'exome, rendant l'identification de variations plus difficile, voire impossible. Ce biais provient d'une part de la technique de séquençage au cours de laquelle l'amplification des régions de faible complexité est plus difficile que le reste du génome ; d'autre part de la capture de ces régions qui est moins efficace en raison d'une baisse de l'hybridation (Clark et al., 2011). L'amélioration de la couverture de ces régions riches en GC se fera avec l'amélioration de ces étapes du séquençage. Plus généralement, les régions de faible complexité, c'est-à-dire des régions de composition nucléotidique biaisée et souvent composées de motifs répétés constituent une limite de l'analyse de données d'exome. Ces segments d'ADN difficiles à aligner sur une séquence de référence présentent des scores d'alignement biaisés car ils peuvent s'aligner à différents endroits de la référence. Leur similarité de séquence empêche l'identification de leur position exacte dans le génome et conduit à des alignements locaux imprécis. Ils ne sont donc pas pris en compte lors du traitement des données de séquençage.

Les pathologies, dites en mosaïque, représentent également un défi pour la détection des variations causales. Elles sont dues à des variations post-zygotiques présentes dans une partie des cellules du corps humain, parfois en faible proportion. Lorsque le pourcentage de lectures possédant la variation est trop faible, cette dernière risque de ne pas être détectée par un pipeline bioinformatique « classique » (Lee et al., 2014a). Dans ce cas, l'augmentation de la profondeur de séquençage, couplée à l'adaptation du pipeline d'analyse de données d'exome, sera nécessaire pour s'adapter à cette spécificité (**Tableau 3**). Ainsi, il est possible d'abaisser le seuil de détection des variations pour atteindre une balance allélique de 10 %, ou de faire appel aux programmes de détection de variations somatiques utilisés dans l'étude de cancers car ils sont spécialisés dans la détection de variations à balance allélique faible (Manheimer et al.,

2018).

L'une des limites inhérentes au séquençage d'exome est sa dépendance aux kits de capture. L'évolution de ces derniers a permis d'améliorer la sensibilité de l'exome. La capture a d'abord été réalisée par des sondes à ADN simple brin. Puis le choix s'est porté sur des sondes à ARN (Agilent) ce qui améliore la force de liaison entre région cible et sonde. Récemment, de nouveaux kits de capture (Twist Biosciences) faisant appel à des sondes à ADN double-brin sont apparus sur le marché. Du fait de la capture dans les 2 sens de lecture, la couverture des régions cibles est améliorée sans avoir besoin d'augmenter la profondeur de séquençage.

L'analyse d'exome ne porte que sur les régions exoniques et les sites d'épissage. Cela ne représente qu'environ 2 % du génome (Mazzarotto et al., 2020). Il s'agit d'une limite soulignée dès le début du séquençage d'exome (Choi et al., 2009). Les 98 % restants comprennent les régions introniques et régulatrices, qui peuvent contenir elles aussi des variations pouvant avoir un effet pathogène et être donc à l'origine de pathologies (Vaz-Drago et al., 2017).

Enfin le dernier niveau de difficultés rencontrées dans l'étude de données de séquençage à haut débit d'exome concerne **l'analyse bioinformatique** elle-même (**Tableau 3**). Parmi les limites bioinformatiques les plus importantes se trouvent la mise à jour des bases de données d'annotation de variations, la détection des CNV et des variations structurales, la mise en évidence des disomies uniparentales, l'identification des STR pathogènes (« Short Tandem Repeat »), la détection des éléments mobiles et l'analyse de l'ADN mitochondrial.

Tout d'abord, l'analyse de données d'exome dépend de la version des bases de données de variations rapportées. Ces bases sont décrites en fonction d'une version du génome donnée. Le pipeline d'analyse utilisé doit donc travailler sur la même version. Par exemple, la référence génomique la plus récente est GRCh38.p13, déployée en février 2019. La base de données gnomAD, utilisée dans de nombreux pipelines, n'est passé de GRCh37 à GRCh38 qu'en octobre de cette même année. Ce délai existant entre les bases de données et les versions du génome de référence ralentit l'amélioration de l'analyse des résultats obtenus par les pipelines bioinformatiques.

Par ailleurs, il existe au sein du génome des modifications de structures chromosomiques de taille supérieure à 1 kb (Feuk et al., 2006) : les variations structurales. Il s'agit d'inversions, de translocations, de Copy Number Variations incluant des délétions et des duplications. Elles peuvent être à l'origine de pathologies rares (Feuk et al., 2006). L'identification des SV est difficile à partir de données de séquençage à haut débit d'exome (Lee et al., 2014a). La détection est rendue possible par des programmes spécifiques à l'exemple de XHMM (Fromer and Purcell, 2014). Mais l'exome représentant environ 2 % du génome, la probabilité que ces événements, et leurs points de cassure, soient présents au sein d'une région exonique, est très faible. Si les inversions et les translocations sont très difficilement détectables, des CNV peuvent être identifiés. La détermination des points de cassure reste cependant imprécise. La démocratisation du séquençage d'exome (GS) permettra de palier cette limite.

Outre la détection des CNV, l'identification des éléments mobiles à partir de données de séquençage à haut débit d'exome représente un défi. Ces éléments, absents de la référence génomique en cas d'insertion récente, ne sont pas aisément détectables avec une méthode dite « classique » d'analyse d'exome car ils présentent des séquences répétées ou génèrent des lectures multimappées. Il existe des programmes capables d'identifier l'insertion d'éléments mobiles à partir de données de séquençage à haut débit (Goerner-Potvin and Bourque, 2018). Néanmoins, comme pour les CNV, la probabilité que l'insertion et donc le point de cassure se situent au niveau des régions géniques séquencées en exome est très faible. La position exacte des points de cassure en dehors des régions exoniques ne peut donc être déterminée qu'avec le séquençage du génome complet. La mise en évidence de l'insertion d'éléments mobiles au sein des régions géniques a fait partie du travail de thèse.

Parmi les difficultés bioinformatiques se trouvent également la détection des STR pathogènes. Créées à la suite d'une erreur de l'ADN polymérase lors de la réplication, ces répétitions en tandem de 1 à 6 nucléotides représentent 3 % du génome humain et 8 % d'entre elles sont situées au sein des régions codantes (Fan and Chu, 2007). Le taux de mutation de ces STR ($\sim 10^{-4}$ mutations par génération par position) est beaucoup plus important que celui des SNV ($\sim 10^{-8}$ mutations par génération par position) (Tang et al., 2017). On dénombre plus de 20 pathologies impliquant un STR, comme l'exemple du syndrome de l'X fragile (OMIM 300624) ou de la dystrophie myotonique (OMIM 160900). Toutes les sections des régions

géniques peuvent être impactées du promoteur à l'extrémité 3'-UTR. Les STR sont difficilement détectables par un alignement classique de short-reads. Leur identification nécessite donc un outil dédié à l'instar de Tredparse (Tang et al., 2017), exSTRa (Tankard et al., 2018), ou ExpansionHunter (Dolzhenko et al., 2019). Si ces outils ont été implémentés pour des données de génome, l'étude de Tankard et al. a montré qu'ils étaient compatibles avec des données d'exome.

Enfin, l'analyse de données issues de la capture indirecte peut s'avérer utile mais difficile. Un des exemples est celui de l'ADN mitochondrial, déjà cité par Lee en 2014 (Lee et al., 2014a). Bien qu'extrait et séquencé en simultané avec l'ADN nucléaire, il reste peu étudié lors des expériences de ES. Certaines pathologies ont pour origine une variation dans l'ADN mitochondrial. Ces maladies mitochondriales constituent un groupe de maladies très hétérogènes ayant pour conséquence une anomalie de la chaîne respiratoire mitochondriale. L'incidence est estimée à 1/5000 (Bannwarth et al., 2013). Ces pathologies ont une expression très hétérogène, aussi bien dans la présentation clinique que dans l'âge d'apparition des symptômes. En effet, les maladies mitochondriales peuvent apparaître à n'importe quel âge, de la période néonatale à l'âge adulte, et l'atteinte peut être de sévérité variable. Devant l'hétérogénéité des maladies mitochondriales, l'identification de variations impliquées dans le phénotype des patients nécessite une analyse complète du génome mitochondrial.

Limites de l’ES	Solutions
Clinico-biologiques	
Connaissances actuelles	Recherche Partage de données Réanalyse des exomes
Techniques	
Régions riches en GC / Traitement des régions de faible complexité	Amélioration de l’amplification PCR Séquençage sans PCR
Maladies en mosaïque	Adaptation du pipeline Augmentation de la profondeur de séquençage Diminution du seuil de détection des variations
Détection des variations exoniques et d’épissage uniquement	Séquençage des régions non codantes (GS)
Kits de capture	Amélioration des sondes
Bioinformatiques	
Génome de référence imparfait	Mise à jour des bases de données et du génome
Détection des CNV/SV	Développement d’outils spécifiques Amélioration de la détection par GS
Détection des STR	
Détection des EM	
ADNmt variations ponctuelles	Développement d’un pipeline dédié
ADNmt CNV	Développement d’outils spécifiques

Tableau 3 : Principales limites de l’analyse et de l’interprétation des données de ES

En plus des séquences de l'ADN mitochondrial, des données « off-target » peuvent aussi donner accès à des régions introniques ou intergéniques. Ainsi il est possible par exemple d'identifier des SNP (Samuels et al., 2013), des CNV non exoniques (Laver et al., 2019) ou des variations somatiques (Lelieveld et al., 2016) à partir de ces données.

III.2- L’essor du séquençage à haut débit du génome

L’arrivée du séquençage à haut débit de génome a révolutionné l’analyse de données issues de l’ADN de patients. Il a été montré que, bien qu’il soit plus coûteux (**Fig. 24**), il est plus performant avec un taux diagnostique attendu de l’ordre de 60 % (Gilissen et al., 2014). En effet, il a permis de palier certaines difficultés rencontrées lors de l’analyse de données d’exome.

Par exemple, au niveau technique, le séquençage du génome a apporté une amélioration dans le traitement des régions riches en GC. En effet, le séquençage à haut débit de génome ne fait pas appel à une amplification PCR, étape qui peut conduire à des défauts d’amplification ou l’apparition de duplicats qui compliquent l’analyse des données issues du séquençage d’exome.

Par ailleurs, les variations introniques des régions régulatrices et intergéniques, non disponibles en exome et pouvant être à l’origine de maladies rares, sont identifiables grâce au séquençage du génome. Leur détection fait alors appel à un « variant caller », c’est-à-dire un programme de détection des variations à partir de données alignées (fichier BAM), programme identique à celui utilisé pour les données de l’exome comme par exemple HaplotypeCaller de GATK (McKenna et al., 2010). La mise en évidence de ces variations est considérée comme une avancée dans l’identification des variations causales impliquées dans des pathologies génétiques rares. Néanmoins, à l’heure actuelle, le nombre de variations dans ces régions est massif, et leur interprétation extrêmement complexe.

Parallèlement, le séquençage de régions non exoniques permet d’obtenir les coordonnées des CNV identifiés par l’analyse bioinformatique, et de détecter le reste des SV. Les SV se forment par cassures chromosomiques, qui ont souvent lieu dans les régions non codantes intergéniques ou introniques, des régions répétées difficiles à aligner sur l’ADN de référence. De plus, certains SV sont le résultat de plusieurs événements successifs, ce qui complexifie d’autant plus l’étape d’alignement des séquences. En effet, les lectures ne s’alignent alors que partiellement sur le génome de référence. Des algorithmes différents de ceux de la détection de variations ponctuelles sont donc utilisés, comme par exemple LUMPY

(Layer et al., 2014) et Control-FREEC (Boeva et al., 2012). L'identification précise des points de cassure permet également d'améliorer la mise en évidence des éléments mobiles en dehors des régions exoniques. Néanmoins, la détection des CNV reste imparfaite et fait l'objet du développement de nombreux outils et méthodes différentes (Zhang et al., 2019).

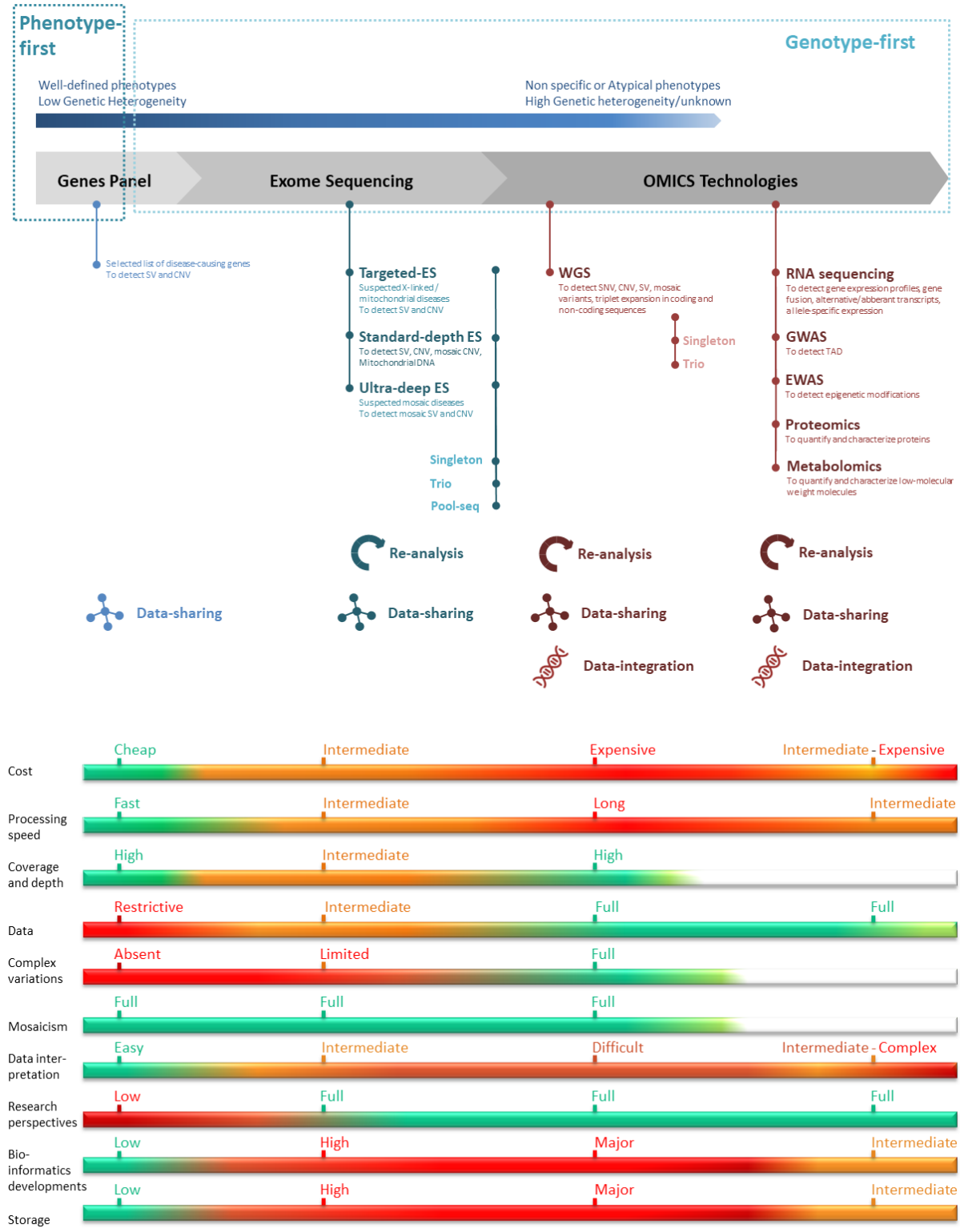


Figure 24 : Avantages et restrictions des différentes technologies d'étude de pathologies génétiques rares

Plusieurs technologies sont disponibles pour identifier la/les variation(s) responsable(s) du phénotype des patients. Pour chaque étude le choix de la stratégie (combinaison de technologies) dépend du coût, du taux de couverture nécessaire, etc. (d'après (Bruel et al., 2020)).

Néanmoins, la technologie de séquençage à haut débit du génome n'est que très peu déployée en routine. En effet, elle reste actuellement plus coûteuse, et la génération d'une quantité massive de données nécessite un stockage et un temps d'analyse important (**Fig. 24**). La réduction du temps d'analyse constitue un véritable défi. Outre les avancées technologiques en matière d'informatique grâce aux nouveaux matériaux d'analyse (GPU, FPGA), l'utilisation d'algorithmes de plus faible complexité, de scripts et de langages optimisés permettront de réduire le temps d'analyse. De plus, toutes les informations provenant du séquençage d'exome n'ont pas été extraites (**Fig. 25**). Il existe, au sein des données brutes d'exome, des données non exploitées, qui peuvent être encore étudiées par l'amélioration des connaissances, ainsi que des analyses bioinformatiques supplémentaires.

Si le séquençage haut débit de génome permet d'expliquer une partie supplémentaire des patients sans diagnostic, certaines situations resteront inexpliquées, et nécessiteront une amélioration des connaissances et des concepts. Citons en particulier les situations de pénétrance incomplète, où différents allèles agissent de façon synergique les uns sur les autres et peuvent avoir des effets modificateurs sur l'expression et/ou la pénétrance de la pathologie rendant son diagnostic plus difficile à poser, jusqu'à l'hérédité oligogénique (Badano and Katsanis, 2002). Nous connaissons également depuis longue date d'autres mécanismes, telles que les modifications épigénétiques, qui nécessiteront d'autres types d'analyse. La place des Omics (transcriptome, protéome, epigénome, métabolome, etc.) est à définir (**Fig. 25**).

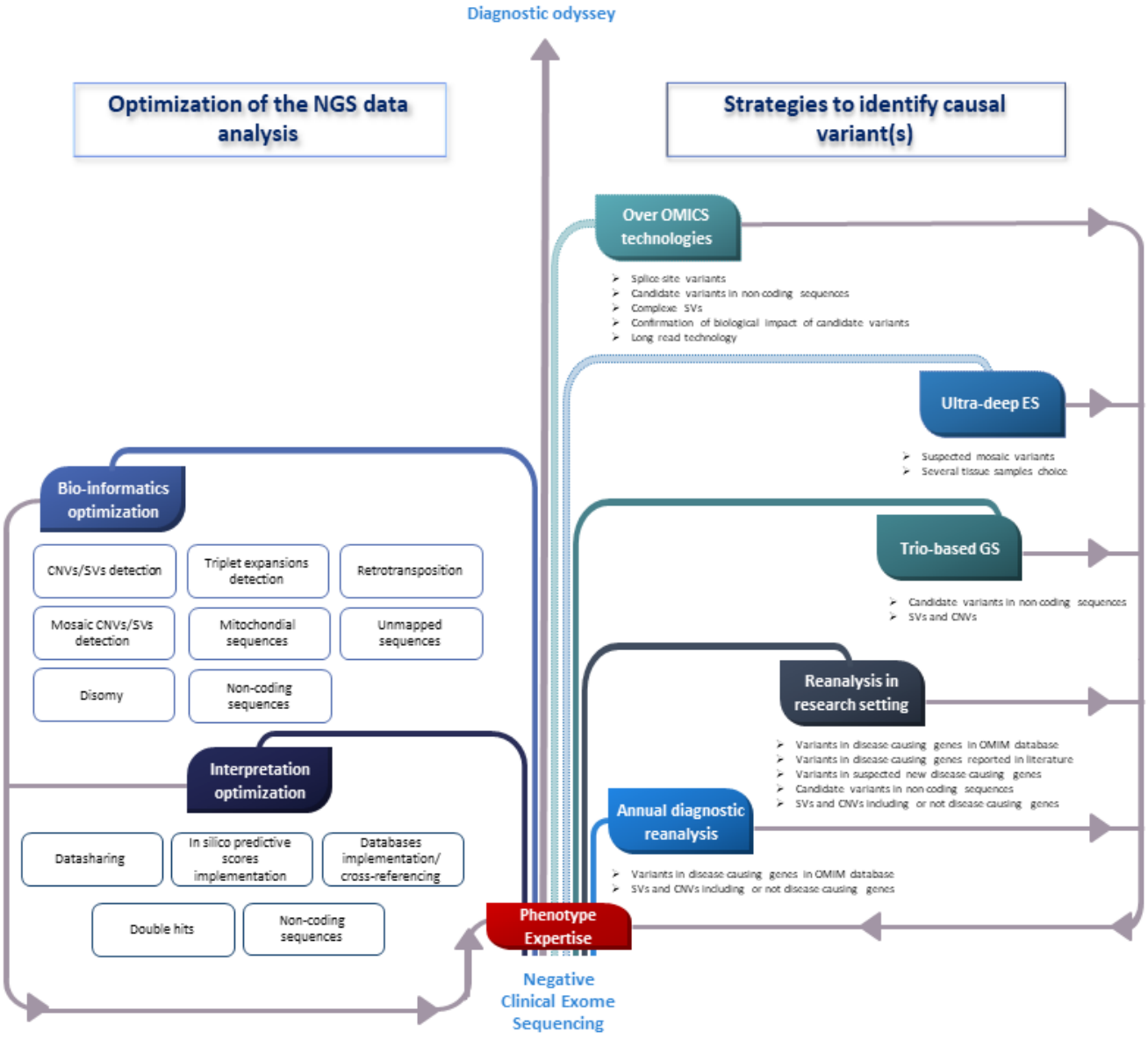


Figure 25 : Stratégies de réduction de l'odyssée diagnostique des patients avec ES diagnostique négatif

La réduction de l'errance diagnostique passe tout d'abord par l'optimisation de l'analyse des données de ES préexistantes. La réanalyse annuelle, la lecture en recherche et le séquençage des parents (analyse en trio) sont des stratégies qui permettent également d'augmenter le taux de diagnostics. Enfin, de nouvelles approches Omics sont nécessaires pour identifier les variations complexes, non exoniques, confirmer la pathogénicité, etc. (d'après (Bruel et al., 2020)).

IV- ANOMALIES DU DÉVELOPPEMENT ET DÉFICIENCE INTELLECTUELLE

IV.1- Définition et épidémiologie

Les anomalies du développement (AD) surviennent au cours du développement embryonnaire ou fœtal et correspondent à des atteintes dans la formation des organes ou des tissus. Elles peuvent être associées ou non à une déficience intellectuelle (DI). La DI est la limitation des capacités cognitives et adaptatives débutant pendant la période du développement. Elle est classée en 4 catégories en fonction du quotient intellectuel (QI) du patient : légère (QI entre 50 et 70), modérée (QI de 40-50), sévère (QI de 20-40) et profonde (QI $\leq$ 20). Sa prévalence dans la population générale est estimée à 3 % des nouveaux-nés (Aicardi, 1998). Cette affection est hétérogène d'un point de vue clinique et étiologique. Environ 15 % des cas seraient d'origine environnementale, 45 % d'origine génétique et 40 % de cause indéterminée. Il est néanmoins probable que les causes génétiques représentent une grande part des causes dites indéterminées. Sa présentation est soit isolée, soit associée à des anomalies du développement (DI syndromique).

Les causes génétiques de la déficience intellectuelle sont elles-mêmes diverses. Dix pour cent des DI d'origine génétique sont dues à des anomalies chromosomiques. Des micro-remaniements génomiques peuvent également être impliqués. Les variations ponctuelles sont les évènements génétiques les plus fréquents. La transmission de cette affection peut donc être dominante, récessive ou liée à l'X. Chaque anomalie génétique représente moins de 1 % des étiologies, exceptés la trisomie 21 et le syndrome de l'X-Fragile qui sont les premières causes de DI (Friedman et al., 2006). Les différentes étiologies de DI sont donc considérées comme des maladies rares. La cause génétique d'environ 800 sous-groupes de DI, isolées ou syndromiques, a été identifiée. Ces dernières années, l'essor du séquençage à haut débit a permis de décrire un nombre croissant de gènes associés à une pathologie humaine.

IV.2- L’essor du séquençage à haut débit de l’exome dans les AD/DI

Jusqu’au milieu des années 2000, l’identification de variations pathogènes causales ou de nouveaux gènes impliqués dans des pathologies génétiques rares faisait appel à un séquençage ciblé par la méthode de Sanger. Cette technique ne permettait le séquençage que d’un nombre limité de gènes, la rendant incompatible avec le bilan d’étiologique d’AD avec ou sans DI. C’est l’application de la méthode de capture à grande échelle par hybridation, qui a permis de séquencer l’ensemble des régions codantes du génome, dit exome.

Ainsi en 2009, Ng et son équipe ont décrit la capture et le séquençage de l’exome de 8 individus, précédemment caractérisés, issus de la base de données HapMap et de 4 patients atteints du syndrome de Freeman-Sheldon (OMIM 193700), pathologie à transmission autosomique dominante, habituellement sporadique (Ng et al., 2009). La comparaison des résultats de l’exome des 8 premiers individus avec les données d’HapMap a mis en évidence une concordance de plus de 99 %. Forts des excellents résultats de sensibilité et spécificité, ils ont appliqué cette méthode aux 4 patients afin d’identifier la cause génétique de leur pathologie commune. Le seul gène au sein duquel les 4 patients possédaient au moins une variation non synonyme était le gène *MYH3* (OMIM 160720), rapporté quelques années plus tôt comme étant responsable du syndrome de Freeman-Sheldon (Toydemir et al., 2006). Ng et al. ont donc présenté la première étude de faisabilité et démontré l’utilité du séquençage d’exome, sans analyse de liaison ni approche par gènes candidats, pour la détection de variations causales pour identifier des gènes responsables d’anomalies du développement.

La même année, Choi et son équipe ont confirmé la faisabilité et l’utilité de cette méthode et ont présenté son application possible dans le diagnostic génétique (Choi et al., 2009). L’étude présente l’analyse de l’exome d’un patient issu d’une union consanguine et suspecté d’être atteint du syndrome de Bartter (OMIM 241200). Les auteurs ont identifié une variation faux-sens homozygote dans le gène SLC26A3 (OMIM 126650). Des variations récessives perte de fonction dans ce gène, c’est-à-dire à l’origine de la perte partielle voire totale d’expression ou d’activité protéique, avaient déjà été décrites comme responsables d’une autre pathologie : la diarrhée chlorée congénitale (OMIM 214700). Ce diagnostic clinique inattendu a par la suite été confirmé par des analyses biochimiques. Il s’agit du premier cas décrit démontrant l’intérêt du séquençage d’exome pour identifier des diagnostics différentiels

non suspectés cliniquement. Cette approche est nommée « genotype-first », c'est-à-dire guidée par le génotype et non par la clinique.

En 2010, Ng et son équipe ont présenté la première application du séquençage d'exome dans l'identification de la cause génétique d'une maladie mendélienne rare (Ng et al., 2010). Il s'agissait de 4 patients atteints du syndrome Miller (OMIM 263750), pathologie à transmission autosomique récessive, qui était alors seulement caractérisé phénotypiquement. Les auteurs ont appliqué une approche « phenotype-first », visant à identifier un gène candidat à partir d'une cohorte d'individus regroupés pour leur phénotype commun. Ils ont ainsi identifié le gène candidat *DHODH* (OMIM 126064) responsable de ce syndrome, depuis considéré comme responsable du syndrome Miller dans la base de données OMIM. Ng et al. ont ainsi mis démontré l'intérêt de l'analyse de données de séquençage d'exome dans l'avancée des connaissances en matière d'identification de gènes responsables d'AD et plus largement des maladies rares monogéniques, y compris dans les formes sporadiques, jusque-là résistantes aux anciennes technologies.

Toujours en 2010, l'équipe de Vissers a utilisé le séquençage d'exome en trio (cas index + parents) pour identifier la cause génétique de la DI non syndromique de 10 patients (2 filles et 8 garçons) sans diagnostic étiologique (Vissers, 2010). Cette étude a permis de mettre en évidence 10 variations *de novo*, menant à un diagnostic confirmé ou probable chez 6 patients (soit 60 % de diagnostic), aussi bien au sein de gènes déjà connus dans des pathologies neuro-développementales que des gènes à fonction inconnue. Les auteurs ont été les premiers à utiliser une approche en trio pour identifier des variations causales *de novo* et des gènes candidats dans des pathologies complexes et variées.

Depuis, de nombreuses études faisant appel au séquençage d'exome dans les AD/DI ont été rapportées : l'utilisation du séquençage à haut débit d'exome dans le diagnostic et l'identification de gènes impliqués dans les anomalies du développement et/ou la DI ne cesse de croître (**Fig. 26**).

La première analyse d'une cohorte de grande taille (136 familles) a été effectuée en 2011 par l'équipe de Najmabadi (Najmabadi et al., 2011). Les auteurs cherchaient à améliorer le rendement de l'analyse d'exome de patients atteints de DI à transmission récessive (10 % des cas). En se concentrant sur des familles consanguines, ils ont ainsi mis en évidence, par séquençage d'exome solo, la cause génétique de la pathologie récessive chez environ 60 % des

patients.

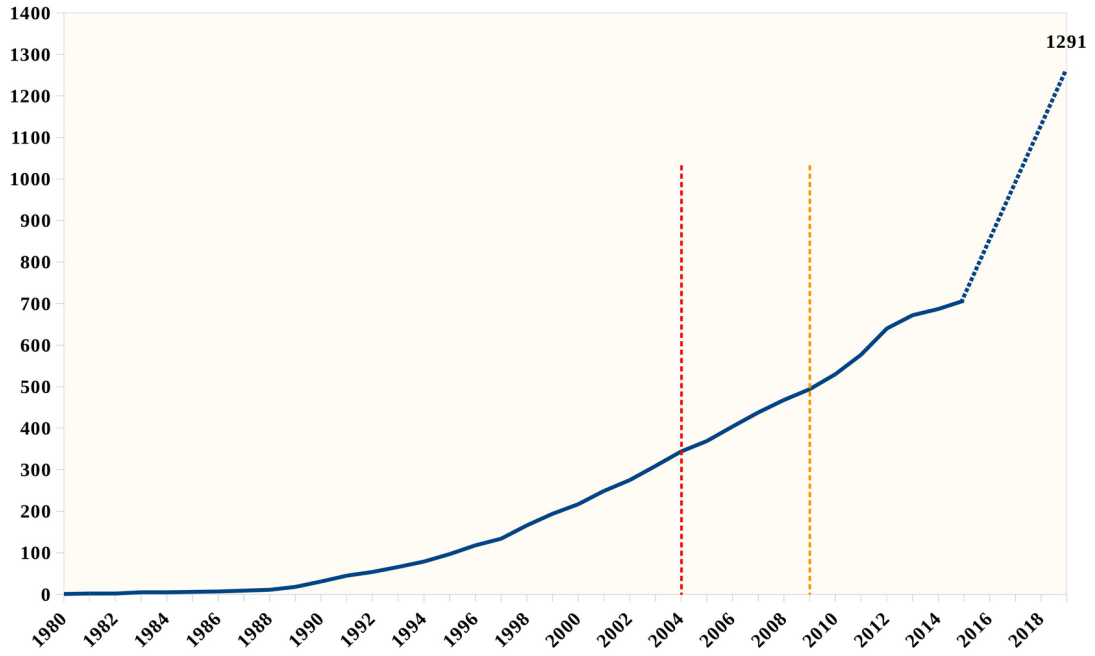


Figure 26 : Nombre de gènes identifiés dans la DI isolée ou syndromique depuis 1980 *L'introduction de la puce à ADN en routine est signalée en rouge, celle de l'ES en orange. Le séquençage à haut débit a permis d'intensifier d'accroître de manière significative le nombre de gènes identifiés chaque année dans ces pathologies génétiques (adapté de (Vissers et al., 2016) et (Kochinke et al., 2016) en date du 4 décembre 2019).*

L'année suivante, l'équipe de Need a souligné l'importance du séquençage d'exome dans l'identification de la cause génétique d'anomalies du développement ou de DI sans diagnostic étiologique (Need et al., 2012). Ils ont séquencé, en trio, l'exome de 12 patients aux phénotypes variés pour lesquels d'autres tests génétiques réalisés n'avaient pas été contributifs. Avec un taux diagnostique de 50 %, ils ont souligné l'importance de cette méthode d'analyse en plein essor dans l'extension de spectres phénotypiques et la diminution de l'errance diagnostique. Cette étude est également revenue sur les limites de l'analyse d'exome. Ainsi les auteurs ont rappelé qu'il existait des exons non couverts, que les régions régulatrices non-codantes ne sont pas séquencées et que les SV sont difficilement identifiables via le séquençage d'exome. Ils proposaient déjà le GS comme solution pour améliorer l'identification des variations causales. Enfin, ils soulignaient l'importance de l'existence de bases de données alliant variations causales et phénotypes ; ainsi que l'étude de l'impact des variations candidates sur l'expression et l'épissage des gènes concernés pour confirmer le caractère pathogène de ces

dernières. Toujours en 2012, l'équipe de De Ligt s'est intéressée au diagnostic de DI sévère (QI <50) sporadique et restée inexpliquée chez 100 patients par séquençage d'exome en trio des cas index et des parents sans consanguinité (de Ligt et al., 2012). La séquence des gènes candidats ainsi identifiés a été ensuite réanalysée chez 765 autres patients également atteints de DI. Les auteurs se sont concentrés sur les variations *de novo* et ont obtenu un taux diagnostique de 16 %. Simultanément, l'équipe de Rauch a travaillé sur une cohorte de 20 contrôles et de 51 patients atteints de DI sévère (QI <60) non syndromique (Rauch et al., 2012). En se concentrant eux aussi sur les variations *de novo* détectées en trio, ils ont obtenu un taux diagnostique de 51 %. Après les travaux de Najmabadi (2011) sur l'importance du séquençage d'exome pour le diagnostic de la DI en cas de consanguinité, les équipes de De Ligt et de Rauch ont montré le succès de cette technologie en première intention dans le diagnostic de la DI sévère et inexpliquée en général.

L'année suivante, Yang et son équipe ont fourni la première estimation d'un taux diagnostique du séquençage d'exome, sans biais d'échantillonnage ni de consanguinité, dans la DI avec ou sans anomalie du développement (Yang et al., 2013). En étudiant en solo 250 patients dont 80 % avec des atteintes neurocognitives suspectées d'être d'origine génétique, ils ont obtenu un taux diagnostique de 25 %. Ils ont justifié cette valeur, par rapport à celle des études précédentes, par l'hétérogénéité des présentations cliniques des patients. Les auteurs ont également présenté des améliorations possibles du taux diagnostique par la détection des CNV et l'amélioration des connaissances. L'existence de variations dans les régions non-codantes et d'exons mal couverts en exome a également été soulevée. Les auteurs ont proposé une solution potentielle en recourant au séquençage du génome pour palier ces difficultés. Mais ils rappellent que le séquençage du génome, en 2013, était beaucoup plus coûteux que celui de l'exome avec une profondeur de séquençage moins importante. Ces limites rendaient pour eux le génome moins intéressant que l'exome à cette époque. Enfin, Yang et son équipe ont pointé du doigt l'existence de « multi-hits », c'est-à-dire des patients porteurs d'un minimum de deux variations causales dans deux gènes distincts qui expliqueraient la superposition des signes cliniques.

La première étude comparant le taux diagnostique de l'analyse en trio à celui du solo chez des patients avec des présentations cliniques variées a été effectuée en 2014 par Lee et son équipe (Lee et al., 2014a). Toutes pathologies mendéliennes confondues, les auteurs ont obtenu

un taux diagnostique en solo de 22 % et de 31 % en trio. Pour les patients atteints de retard du développement (37 % de la cohorte), le taux diagnostique était de 28 %. Ainsi les auteurs ont souligné que l'analyse en trio permettait de diminuer l'errance diagnostique des patients, notamment ceux avec atteintes neurocognitives, par la détection facilitée de variations causales *de novo* ou hétérozygotes composites. Ils ont également rappelé les limites du séquençage d'exome qui devront être améliorées pour augmenter le taux diagnostique : état des connaissances, fréquence allélique des variations rares d'individus contrôles, détection des CNV, mise en évidence des disomies uniparentales, détection des variations somatiques en mosaïque, détection des répétitions de triplets et analyse de l'ADN mitochondrial. La détection des disomies uniparentales, des variations en mosaïque et l'analyse de l'ADN mitochondrial n'ont pas été réalisées de manière automatique lors de cette étude, mais les auteurs ont identifié des cas manuellement. Toujours en 2014, Farwell et son équipe ont effectué la même comparaison entre trio et solo sur une cohorte de 500 patients dont 64 % étaient atteints de DI et/ou retard du développement (Farwell et al., 2015). Les taux diagnostiques de 37 % en trio et de 21 % en solo et la moyenne de 30 % sont du même ordre de grandeur que ceux obtenus par Lee et son équipe. Ils ont eux aussi rappelé l'apport du séquençage d'exome dans le diagnostic et dans la recherche de nouveaux gènes impliqués en pathologies humaines.

Parmi les études qui ont suivi, il peut être noté celles de Yang et al. (Yang et al., 2014) et de Wright et al. (Wright et al., 2015), réalisées toutes deux en 2014 sur des cohortes de plus de 1000 patients. Pour la première, les auteurs ont analysé 2000 exomes en solo d'une cohorte de patients avec 26 % d'atteintes neurocognitives isolées et 55 % d'atteintes neurocognitives syndromiques. Ils ont obtenu des taux diagnostiques de 27 % et 25 % respectivement. Pour la deuxième, effectuée sur 1133 exomes en trio de patients atteints d'anomalies du développement dont 87 % avec DI, le taux diagnostique était également de 27 %.

Ainsi, le taux diagnostique de la DI qui stagnait aux alentours de 10 % jusque dans les années 2000, a vu sa valeur démultipliée par l'introduction du séquençage d'exome et l'amélioration des connaissances pour atteindre 30 % à 45% aujourd'hui (Bruel et al., 2019a), voire 70 % pour la DI sévère dans des populations consanguines (**Fig. 27**). Le séquençage d'exome n'identifie donc pas la cause de la maladie chez environ deux tiers des patients car certaines limites existent à plusieurs niveaux de cette étude : clinico-biologique, moléculaire et bioinformatique. Ces limites conduisent ainsi à des difficultés d'analyse.

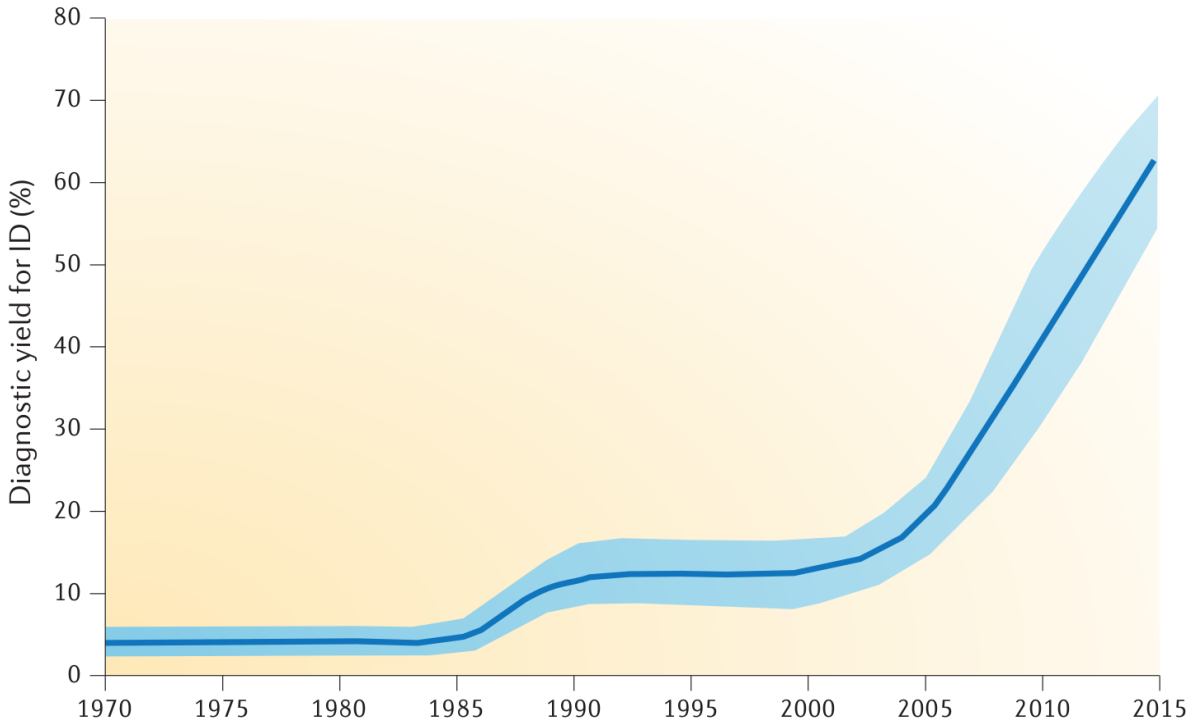


Figure 27 : Evolution du taux diagnostique de la DI

Dans les années 70 le diagnostic était effectué par caryotype dont le taux diagnostique des DI variait entre 3 et 6,5 %. La première amélioration a eu lieu notamment par l'introduction du séquençage Sanger en routine (début des années 90) améliorant le taux diagnostique de 6 à 10 %. L'apparition des puces à ADN et surtout du séquençage haut débit d'exome ont permis d'augmenter le taux diagnostique successivement de 15-23 % et 24-33 %. Les premières études du séquençage du génome complet ont montré qu'il était possible d'ajouter environ 26 % de diagnostics supplémentaires. Au total le taux diagnostique de la DI peut atteint 55 à 70 % (d'après (Vissers et al., 2016)).

OBJECTIFS DU TRAVAIL DE THÈSE

Le but du travail de thèse a été d'extraire des informations supplémentaires à partir des données de ES afin de répondre à certains défis posés par l'analyse de ces données pour l'identification de nouveaux mécanismes moléculaires impliqués dans des maladies génétiques rares. Plusieurs pistes ont ainsi été explorées.

- Le premier axe a consisté en une réanalyse recherche de données de patients ayant bénéficié d'un ES au laboratoire CERBA (thèse CIFRE) dont la lecture diagnostique était négative (article 1).
- Le deuxième axe a consisté en la mise au point d'un pipeline bioinformatique pour extraire les données du génome mitochondrial à partir des données de ES à partir de la collection GAD d'exomes de patients sans diagnostic (article 2).
- Le troisième axe a consisté en la mise en place d'un pipeline bioinformatique d'identification des éléments mobiles au sein des données d'exome, à partir de la collection GAD d'exomes de patients sans diagnostic (article 3 en cours d'écriture).

PREMIÈRE PARTIE : INTÉRÊT DE LA RÉANALYSE RECHERCHE DES DONNÉES DE SÉQUENÇAGE D'EXOME DANS LES ANOMALIES DU DÉVELOPPEMENT ET DÉFICIENCE INTELLECTUELLE

I- INTRODUCTION

La réanalyse des exomes diagnostiques négatifs se basent à la fois sur l'avancée des connaissances et sur la recherche d'arguments pour impliquer de nouveaux gènes en pathologie humaine.

L'augmentation du taux diagnostique passe en premier lieu par la réanalyse des données en tenant compte de la mise à jour des bases de données. En effet, de nouveaux gènes sont régulièrement impliqués en pathologies humaines (OMIM - Online Mendelian Inheritance in Man., 1996). Plusieurs études ont mis en évidence une augmentation du taux diagnostique allant de 7 à 10% avec cette approche de relecture dans un cadre diagnostique (Wenger et al., 2017 ; Nambot et al., 2018 ; Wright et al., 2018 ; Bruel et al., 2019a ; Al-Nabhani et al., 2018). Il existe également des cas de variations initialement non retenues dans des gènes connus mais qui se sont finalement avérées causales (Nambot et al., 2018). Al-Nabhani et al., évoque par ailleurs l'impact des seuils utilisés pour filtrer les variations. Ainsi, la réanalyse leur a permis d'identifier la cause génétique chez 1 patient supplémentaire (+ 2%) avec une variation au sein d'un gène déjà connu en pathologie lors de la première analyse (Al-Nabhani et al., 2018). L'importance d'une analyse des parents voire d'autres membres de la famille du cas index est également rappelée dans plusieurs études (Farwell et al., 2015 ; Eldomery et al., 2017). Cette approche permet aussi de diminuer le temps d'analyse et de faciliter l'identification de nouveaux gènes.

Lorsque la réanalyse diagnostique ne permet pas d'identifier la cause de la pathologie il est possible d'appliquer une approche de recherche. Certaines équipes décident ainsi d'analyser les variations présentes au sein de gènes non décrits en pathologie humaine. Cette approche dans un cadre de recherche permet d'une part d'améliorer les connaissances sur la génétique des maladies rares et d'autre part d'augmenter (voire doubler) le taux diagnostique d'une réanalyse d'exome. Sur une cohorte de 50 patients, Al-Nabhani et al. ont identifié un nouveau gène augmentant le taux diagnostique de 2 % (Al-Nabhani et al., 2018). Nambot et al. et Bruel et al. ont accru le nombre de diagnostics positifs de presque 8 %. (Nambot et al., 2018 ; Bruel et al., 2019a).

Un des arguments permettant l'implication d'un nouveau gène dans une pathologie est la récurrence des cas. Dans le cas des pathologies génétiques rares, la collaboration entre équipes est donc cruciale. Plusieurs études ont ainsi montré que le partage de données est essentiel pour augmenter le taux de résolution de l'exome (Eldomery et al., 2017 ; Nambot et al., 2018 ; Bruel et al., 2019a ; 2019b)

La réanalyse d'exome permettant d'améliorer le diagnostic des patients de 10-15 % et de faire avancer les connaissances quant aux mécanismes physiopathologiques des maladies génétiques rares, nous l'avons appliqué à la cohorte des exomes diagnostiques du Laboratoire Cerba. En effet, cette thèse est un thèse CIFRE, c'est-à-dire un partenariat entre un laboratoire privé (laboratoire Cerba) et une équipe de recherche académique (équipe GAD à Dijon). Il s'agit ici d'appliquer l'expertise de l'équipe GAD dans la réanalyse recherche d'exome au données de ES du laboratoire Cerba qui réalise la lecture diagnostique.

II- MATÉRIEL ET MÉTHODES

II.1- Cohorte de patients

La cohorte des 172 individus séquencés et analysés par le laboratoire Cerba était constituée de 127 cas index (dont 5 paires de frères/sœurs atteints), 17 pères, 19 mères, 6 pools parentaux et 3 frères/sœurs non atteints. Le taux diagnostique du séquençage haut débit de l'exome était de 30 %. Les 80 cas index négatifs ont été réanalysés en diagnostic puis en recherche dans le cadre de la thèse CIFRE.

Un séquençage à haut débit d'exome en solo ou trio à partir des échantillons ADN avait été réalisé chez chaque individu. La capture avait été effectuée avec le kit Clinical Research Exome V2 (Agilent). Le séquençage avait été réalisée sur un ABI Prism avec une lecture en paired-end et des lectures de 100pb.

II.2- Traitement et alignement des données brutes de séquençage d'exome dans un but de réanalyse

Le pipeline de réanalyse des données d'exome utilisé par le Laboratoire Cerba a été mis au point par l'équipe GAD en suivant les recommandations du Broad Institute (Van der Auwera et al., 2013).

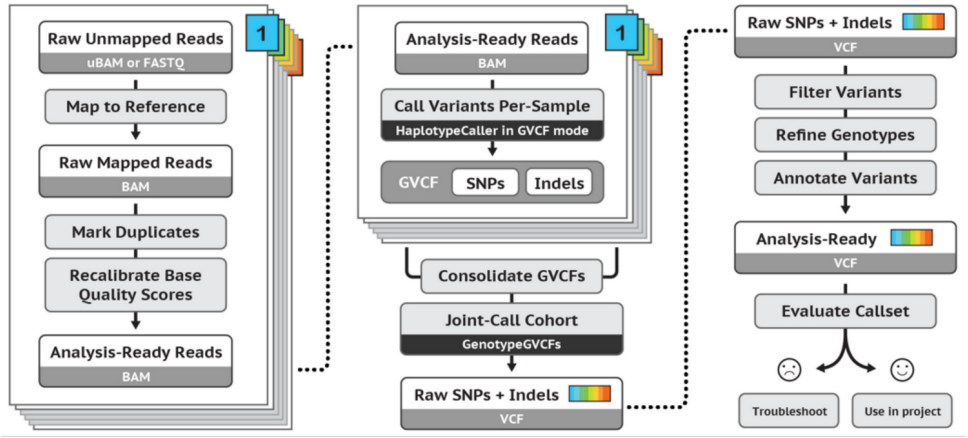


Figure 28 : Pipeline d'analyse de données d'ES conseillé par le Broad Institute
Le pipeline comporte 3 phases. La première consiste en l'obtention d'un fichier d'alignement BAM « propre » à partir des données brutes de séquençage. Vient ensuite la détection des variations (« variant calling »). Enfin les variations brutes sont filtrées et annotées pour faciliter l'interprétation de l'exome (d'après (Broad Institute)).

La qualité des séquences obtenues sous forme de fichiers fastq est analysée avec le programme FastQC (v.0.11.4) (Andrews, 2010). Puis le traitement avec Trimmomatic (v.0.35) (Bolger et al., 2014) permet d'améliorer la qualité des séquences analysées. Les données générées sont ensuite alignées sur un génome de référence GRCh37/hg19 à l'aide de l'outil Burrows-Wheeler Aligner (BWA, (Li and Durbin, 2009)) (v.0.7.15). Le fichier SAM ainsi généré est trié et converti en fichier BAM grâce à Samtools (Li et al., 2009) (v.1.2). Les duplicats PCR sont marqués par le logiciel Picardtools v.2.4.1. Le réalignement des indels puis la recalibration des scores de qualité de base sont réalisés par le logiciel Genome Analysis Toolkit v.3.7 puis v.3.8 (McKenna et al., 2010).

II.3- Analyse des données de l'exome nucléaire

Les variations ponctuelles et les petites indels sont détectées par HaplotypeCaller de GATK v.3.7 puis v.3.8. Les variations ayant passé un premier filtre de qualité (un score de qualité des bases >30 et une qualité d'alignement des lectures >20) sont ensuite annotées avec SnpEff v.4.3T (Cingolani et al., 2012). Les variations retenues sont celles qui affectent la séquence codante ou les sites d'épissage et qui sont qualifiés de rares, c'est à dire qui ont une fréquence inférieure à 1 % (1000 Genomes Project Consortium et al., 2012) dans les bases de données dbSNP Build 151 (Sherry et al., 2001) (http://www.ncbi.nlm.nih.gov/projects/SNP/), ExAC Browser v.0.3.1 (Lek et al., 2016) (http://exac.broadinstitute.org/) et gnomAD v.2.0.2 (Lek et al., 2016) (https://data.broadinstitute.org/gnomAD/release-170228/exomes/vcf/). Les informations issues des bases de données OMIM, ACMG, ClinVar et COSMIC sont également ajoutées ; ainsi que les scores de prédiction bioinformatique CADD, polyPhen, GERP, grantham, MISZ et pLI. Le pipeline génère ainsi un rapport contenant des variations rares et annotées.

De fréquence rare dans la population, les variations candidates ne doivent pas non plus être détectées au sein d'un lot d'une centaine d'individus contrôles et des patients non apparentés du même batch. Les différentes variations sélectionnées sont visualisées sur Integrative Genomics Viewer v.2.4.6 (https://www.broadinstitute.org/igv/home). Deux stratégies de séquençage ont été appliquées : l'exome trio (cas index avec les 2 parents) ou l'exome solo (cas index seul). La réanalyse diagnostique se concentre sur les gènes morbides décrits dans la base de données OMIM en lien avec les bases publiques comme ClinVar. A l'inverse, la lecture recherche concerne les variations présentes dans des gènes non décrits dans une pathologie (**Fig. 29**). Quelles que soient la stratégie de séquençage et le mode de lecture, l'analyse des variants est priorisée en fonction : (i) du mode de transmission suspecté dans la famille (récessif, lié à l'X ou sporadique). L'analyse en trio permet, de plus, l'étude des variations *de novo* chez le patient. (ii) de leurs impacts sur la protéine. Les mutations tronquantes, dont les conséquences protéiques sont considérées comme plus importantes, sont analysées en premier. Puis les variations des sites d'épissage et faux-sens sont étudiées.

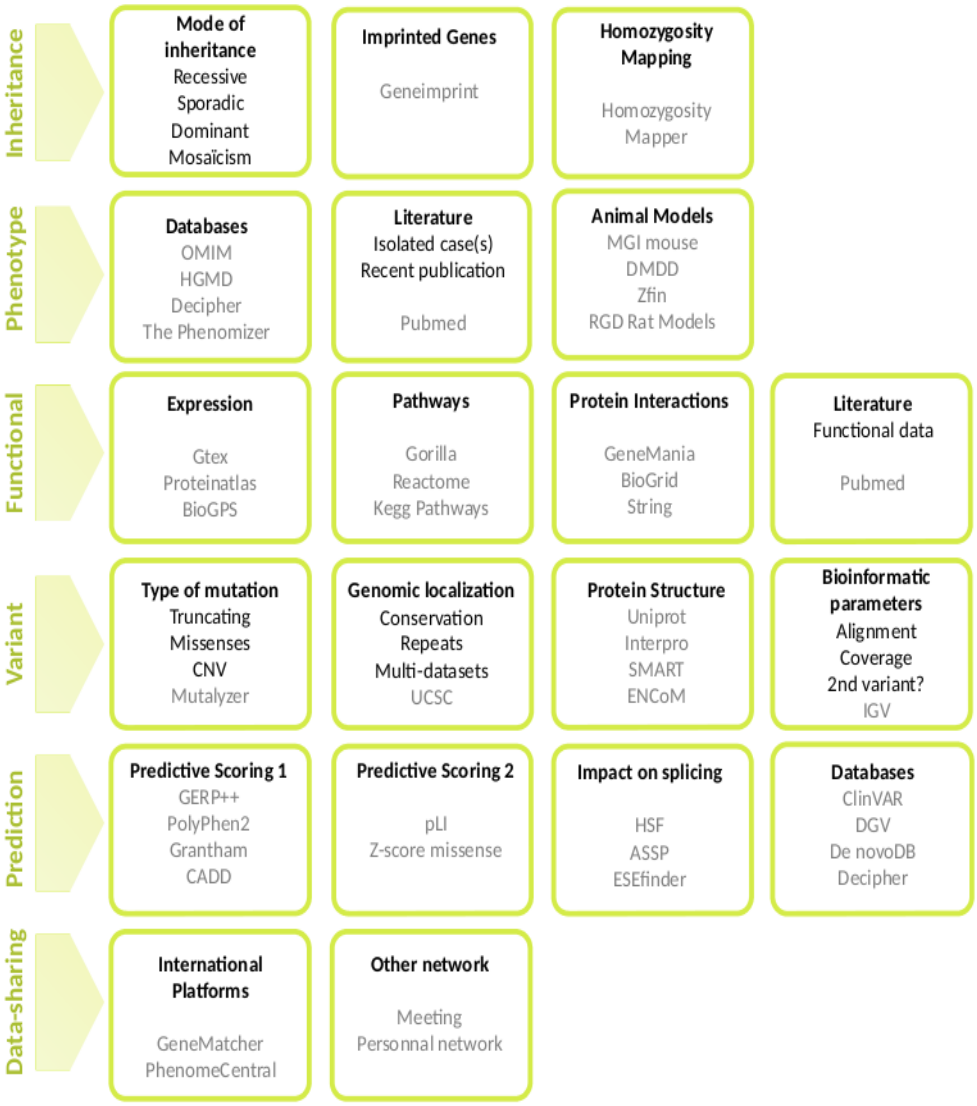


Figure 29 : Stratégie d'analyse des exomes
La réanalyse diagnostique se concentre sur les variations ou les gènes rapportés dans une pathologie humaine. La lecture recherche étudie les gènes non associés à une maladie et réunit le maximum d'informations présentes sur diverses bases de données (Bruel et al., 2019b).

La lecture recherche consiste mettre en évidence un ensemble d'indices qui conduisent à suspecter une variation et son gène comme étant à l'origine du phénotype du patient. La littérature scientifique et les bases de données sont consultées (Pubmed, données d'expression (GTex), modèles animaux (Mouse Genome Informatics), patients rapportés (DECIPHER), etc.). Le phénotype des patients décrits avec une ou plusieurs variations dans le même gène est

comparé avec celui du patient séquencé. Le phénotype des modèles animaux est examiné pour mettre en évidence des similitudes avec les manifestations cliniques du patient ou pour identifier la fonction du gène impacté. Les informations sur les gènes de la même famille que le gène candidat sont également étudiées pour identifier toute interaction, similitude de structure ou co-expression avec des gènes déjà connus en pathologie humaine. Les scores de prédictions bioinformatiques servent à hiérarchiser l'analyse *in silico* et à apporter des informations supplémentaires : PolyPhen (v2.2.2), GERP, Grantham, MISZ, pLI et CADD. Ainsi, les variations tronquantes dans des gènes ayant une pLi à 1 sont étudiées en premier tout comme les variations prédites comme ayant un impact sur l'épissage. Les variations faux-sens sont hiérarchisée avec le score MISZ. La conservation, au cours de l'évolution, de l'acide-aminé muté est également analysée afin de n'étudier que les régions conservées (scores GERP et Grantham). Les scores de prédiction de l'impact des variations sur les protéines (PolyPhen et CADD) sont examinés pour prioriser celles ayant les impacts les plus importants.

II.4- Validation de la variation du gène *OTUD7A* par méthode Sanger

La validation de la variation chr15:g.31819467G>A (NM_130901.2:c.697C>T, p.(Leu233Phe)) du gène *OTUD7A* a été réalisée par séquençage Sanger à l'aide des amorces de PCR sens 5'-TGTTTCTGCCTTCCCCTCAT-3' et antisens 5'-TGGGAAGCCTCAGGAAGAAG-3'. L'ADN génomique d'écouvillons buccaux du frère a été obtenu par extraction au phénol-chloroforme.

II.5- Analyse fonctionnelle du gène *OTUD7A*

L'analyse fonctionnelle a été réalisée par l'équipe du Dr Frédéric Ebstein (Greifswald, Allemagne) dans le cadre d'une collaboration internationale.

II.5.1- Culture cellulaires

Les fibroblastes du patient porteur de la mutation *OTUD7A* ont été cultivés dans du milieu DMEM high-glucose medium (HyClone Thermo Scientific) supplémenté avec 10 % de sérum de veau fœtal et 1 % de réactifs anti-mycoplasmes et antibiotiques (Minerva, Biovalley, France). Les cellules ont été cultivées à 37°C dans une atmosphère humide à 5 % de CO_2.

Les lignées de cellules haploïdes HAP1 parentales et déplétées pour OTUD7A par la méthode de CRISPR/Cas9 ($OTUD7A^-$) ont été fournies par Horizon Discovery. La présence de la variation *knockout* NM_130901.1:c.494_509del, p.Ser165Thrfs*2 a été vérifiée par séquençage Sanger à l'aide des amorces de PCR sens 5'-TAGAATCCTAACTGAAAACTGCCCA-3' et antisens 5'-CTTGGGGCTATCTTTCCTTCTCC-3'. Ces lignées cellulaires haploïdes ont été cultivées dans du milieu Iscove's modified Dulbecco's Medium (IMDM) supplémenté avec 10 % de sérum fœtal de veau et 1 % de Pénicilline/Streptomycine, à 37°C et dans une atmosphère humide à 5 % de CO_2.

II.5.2- Western Blot sur lignées de fibroblastes

Les lignées de fibroblastes contrôles (CTRL1) et mutantes pour OTUD7A ont été lysées dans du tampon TSDG (10 mM Tris pH 7.0, 10 mM NaCl, 25 mM KCl, 1.1 mM MgCl2, 0.1 mM EDTA, 1 mM DTT, 1 mM NaN3, 20 % Glycerol). La quantification protéique a été réalisée selon la méthode de Bradford. Vingt microgrammes d'extrait protéique total ont été chargés sur gels SDS PAGE en conditions non-dénaturantes (3 %-12 %, Invitrogen). Une partie des gels ont été mis en contact avec 0,1 mM de peptide fluorogène (Leu-Leu-Val-Tyr-AMC, Bachem) pendant 30 minutes à 37°C. La fluorescence a été mesurée par le système Fusion FX Imager (Peqlab). Les autres gels ont été transférés sur des membranes PVDF (nitrocellulose). Des anticorps primaires dirigés contre les sous-unités α6 (clone MCP20, Enzo Life Sciences), Rpn5 (clone H3, Santa Cruz Biotechnology Inc.) du protéasome et PA28-α (serum K232/1) ont été mis à hybrider. Puis les anticorps secondaires anti-souris et anti-lapin conjugués à la peroxydase de raifort ont été ajoutés.

L'hybridation a été détectée grâce à la chimioluminescence (Biorad).

Les protéines des lignées de fibroblastes contrôles (CTRL1) et mutés pour OTUD7A ont également été extraites dans du tampon RIPA (RadioImmunoPrecipitation assay buffer, 50 mM Tris pH 7.5, 150 mM NaCl, 2 mM EDTA, 1 % NP40, 0.1 % SDS). Elles ont ensuite été séparées sur gels SDS-Laemmli (10-15 %) et transférées sur membranes PVDF. Les anticorps primaires utilisés pour le Western Blot et dirigés contre les sous-unités Rpt1 (PW8315), Rpt2 (PW0530), Rpt3 (clone TBP7-27), Rpt4 (clone p42-23), Rpt5 (PW8310), Rpt6 (clone p45-110), Rpn10 (clone S5a-18), β1 (clone MCP421), β2 (clone MCP165), β5 (PW8895) du protéasome proviennent de chez Enzo Life Sciences. L'anticorps primaire dirigé contre la sous-unité β5i du protéasome (anti-LMP7, clone A12) provient de chez Santa Cruz Biotechnology. Les anticorps primaires dirigés contre la sous-unité β2i (anti-MECL1, PA5-19146) du protéasome et contre la protéine OTUD7A (PA5-90565) proviennent de chez ThermoFischer Scientific. Les anticorps primaires dirigés contre PA28-β (#2409) et contre l'α-tubuline (clone DM1A) proviennent de chez Cell Signaling Technology. L'anticorps primaire dirigé contre la sous-unité β1i (anti-LMP2, K221) du protéasome provient du stock du laboratoire de l'équipe de Frédéric Ebstein qui a effectué l'analyse fonctionnelle de la protéine OTUD7A. L'hybridation a été détectée grâce à la chimioluminescence (Biorad).

Les protéines des lignées de fibroblastes contrôles (CTRL2, CTRL3 et CTRL4) et mutés pour OTUD7A (c.697C>T) ont également été extraites dans du tampon RIPA (50 mM Tris pH 7.5, 150 mM NaCl, 2 mM EDTA, 1 % NP40, 0.1 % SDS). Elles ont ensuite été séparées sur gels SDS PAGE avant transfert. Les anticorps primaires dirigés contre la protéine OTUD7A, le peptide PA28- α et les sous-unités du protéasome β1, β2, β5, β1i, β5i, Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6 et Rpn5 ont ensuite été mis à hybrider. Les témoins de charge ont été mis en évidence par las anticorps primaires dirigés contre la GAPDH et la β-tubuline.

II.5.3- Western Blot sur lignées haploïdes HAP1

La détection, chez les lignées haploïdes HAP1 parentale et déplétée en OTUD7A (HAP OTUD7A$^-$ des peptides α6, Rpn5 et PA28-α a été réalisée selon le protocole décrit pour les lignées contrôles CTRL1 et mutées pour OTUD7A.
Les lysats cellulaires des lignées HAP1 parentale et OTUD7A$^-$ ont été préparés selon la méthode décrite plus haut pour les lignées contrôles CTRL2, CTRL3 et CTRL4 ainsi que les fibroblastes du patient. L'analyse par Western Blot a été réalisée avec des anticorps dirigés contre OTUD7A, PA28-α, la GAPDH et l'α-tubuline et contre les sous-unités du protéasome α6, β1, β2, β5, β1i, Rpt1, Rpt2, Rpt3, Rpt4, Rpt5 et Rpt6.

II.5.4- Mesure de l'activité chymotrypsin-like

La mesure de l'activité chymotrypsin-like a été réalisée sur une plaque 96 puits. Chaque mesure a été effectuée en quadruplicats. Dix microgrammes de lysats protéiques ont été déposés dans un volume final de 100µL d'une solution à 50 mM NaCl, 50 mM Tris-HCl, pH 7.5, 5 mM MgCl2, 2 mM ATP avec 0.1 mM de peptide substrat Suc-Leu-Leu-Val-Tyr-AMC. La libération d'AMC libre issue de la digestion peptidique a été mesurée toutes les 15 minutes pendant 2h par le lecteur de fluorescence NanoQuant Plate™ (Tecan) à 360/460 nm.

II.5.5- Détection des protéines ubiquitinées via la lysine 48

Le culot protéique, insoluble dans le tampon RIPA, a été resuspendu dans du tampon de lyse contenant de l'urée 8 M, thiourée 2 M et CHAPS 4 %. Le chargement sur gel SDS-PAGE puis le Western Blot ont ensuite été réalisés.

III- RÉSULTATS DE LA LECTURE EN RECHERCHE DES EXOMES NÉGATIFS EN DIAGNOSTIC

La lecture en recherche des exomes négatifs après une réanalyse diagnostique a permis la mise en évidence, dans 2 gènes OMIM non morbides, de 2 variations suspectées d'être responsables de la pathologie génétique des deux patients concernés. Par ailleurs, deux diagnostics supplémentaires ont pu être posés par identification de deux variations dans deux gènes OMIM morbides.

III.1- Identification d'une variation homozygote faux-sens au sein du gène *OTUD7A*

III.1.1- Présentation clinique du patient

Le patient est né à terme après une grossesse normale. Ses mensurations à la naissance étaient dans la moyenne. Il s'agit du deuxième enfant d'un couple d'origine portugaise, d'apparenté lointaine et présentant des difficultés d'apprentissage. Les parents ont arrêté leur scolarité avant le baccalauréat. Durant la consultation de neuropédiatrie, ils ont fait preuve d'une compréhension limitée. Son frère aîné, âgé de 5 ans, est scolarisé en école classique et présente des difficultés d'apprentissage aspécifiques. Néanmoins, il n'a pas été évalué de façon formelle en raison du contexte familial. En dehors de ces éléments notables, la famille ne présente pas d'antécédents particuliers.

La première évaluation du patient a été réalisée à 4 mois suite à l'apparition, décrite par les parents, de mouvements anormaux de type perte de contrôle de la tête pendant plusieurs secondes, associés à une hypotonie. L'IRM cérébrale réalisée à 5 mois était normale. Mais l'électroencéphalogramme (EEG) effectué à 13 mois montrait un motif typique de spasmes infantiles avec hypsarythmie. Ce diagnostic de spasmes infantiles a conduit à la mise en place d'un traitement par vigabatrine et prednisolone. Quatorze jours après l'initiation du traitement,

les spasmes étaient complètement contrôlés. Les différentes EEGs réalisées depuis montrent un tracé de fond anormal continu lors de l'éveil et du sommeil avec des pointes lentes multifocales fréquentes. Le traitement aux corticostéroïdes a pu être progressivement diminué au bout d'un mois. L'IRM de suivi réalisée à 1 an et 3 mois, soit un mois après le début de la prise de vigabatrine, présentait des anomalies aspécifiques. On retrouvait un discret hypersignal T2/Flair des pallidi, d'aspect symétrique, se traduisant par un hypersignal sur la séquence de la diffusion (restriction) ainsi qu'un discret hypersignal de la diffusion de noyau ventro-postéro-latéral du thalamus de façon bilatérale. Il n'y avait pas de pic de lactate sur la séquence de spectroscopie. On observait un élargissement des espaces péri cérébraux, notamment en fronto-pariétal, et un aspect élargi des sillons hémisphériques bilatéraux et discrètement des vallées sylviennes. Le corps calleux était fin, d'aspect dysmorphique. Ces observations étaient compatibles avec celles décrites dans la littérature à la suite d'un traitement par vigabatrin (Dracopoulos et al., 2010). L'analyse du liquide céphalo-rachidien n'a révélé aucune anomalie des acides lactiques et pyruviques. Le traitement par vigabatrin a été arrêté au bout de 4 mois et le patient ne présentait plus de spasmes. Le maintien de la tête a été acquis à 10 mois et le patient a pu se retourner à partir de 12 mois.

Lors de la dernière consultation, à 28 mois, les parents ont précisé que leur fils était capable de s'asseoir sans maintien depuis l'âge de 23 mois. Il n'avait pas acquis la marche mais pouvait se tenir debout quelques secondes. Il n'utilisait que 3 mots : « Papa, Maman, Kévin ». Il ne pouvait ni manger seul, ni pointer du doigt, ni tenir à pleine main. Il était capable de saluer de la main. Comme il a développé la pince pouce-index, il lui était partiellement possible d'attraper, de tenir et de transférer des objets. Il babillait et présentait un contact oculaire et un sourire en réponse à un stimulus social. Il pesait 12,05 kg (-0,5 DS) et son périmètre crânien était de 48,5 cm (-2 DS). Il présentait une hypotonie persistante. Les réflexes ostéotendineux étaient normaux et il n'y avait pas de signes de Babinski ni d'Hoffman, pas de tremblement ou d'hypokinésie. A 28 mois, le patient présentait des absences épileptiques atypiques caractérisées à l'EEG par des pointes-ondes généralisées avec manifestations motrices. Le garçon a été traité avec succès par prises concomitantes de levetiracetam et zonisamide

III.1.2- Réanalyse de l'exome

La réanalyse recherche de l'exome du patient a mis en évidence une variation homozygote faux-sens chr15:g.31819467G>A (NM_130901.2:c.697C>T, p.(Leu233Phe)), au sein du gène *OTUD7A* (OMIM 612024). L'analyse en trio montrait sa présence à l'état hétérozygote chez les parents (**Fig. 30a**), confirmée par le séquençage par méthode Sanger (**Fig. 30b**). L'analyse de ségrégation supplémentaire a montré la présence de cette variation à l'état hétérozygote chez le frère aîné.

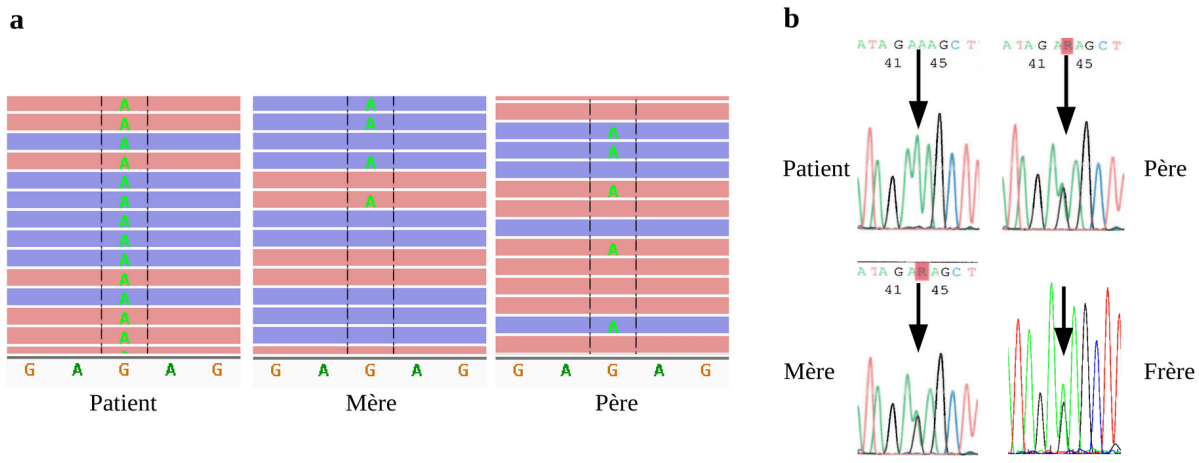


Figure 30 : Variation chr15:g.31819467G>A (NM_130901.2:c.697C>T, p.(Leu233Phe)) au sein du gène OTUD7A
a) Vue IGV de la variation homozygote identifiée par approche d'exome trio. b) Validation Sanger de la variation chez les 4 membres de la famille (sur sang ou écouvillon buccal pour le frère). Le cas index est homozygote tandis que les 3 autres individus sont hétérozygotes pour la variation. La position de la variation est indiquée par la flèche noire.

Cette variation était absente de la base de données gnomAD. Les scores bioinformatiques étaient en faveur d'un effet prédit comme pathogène : CADD=31, GERP=5.6, Polyphen2=1.0, MISZ=5.917 and pLI=0.975. De plus, les scores de prédiction fonctionnelle SIFT et Provean (sur VarSome (Kopanos et al., 2019)) prédisaient un impact négatif de la variation. Cette dernière était située au sein du domaine catalytique « cysteine protease OTU-like ». La leucine affectée faisait partie d'une hélice alpha, conservée au cours de l'évolution et dans la sous-famille des protéines OTU A20-like (**Fig. 31**).

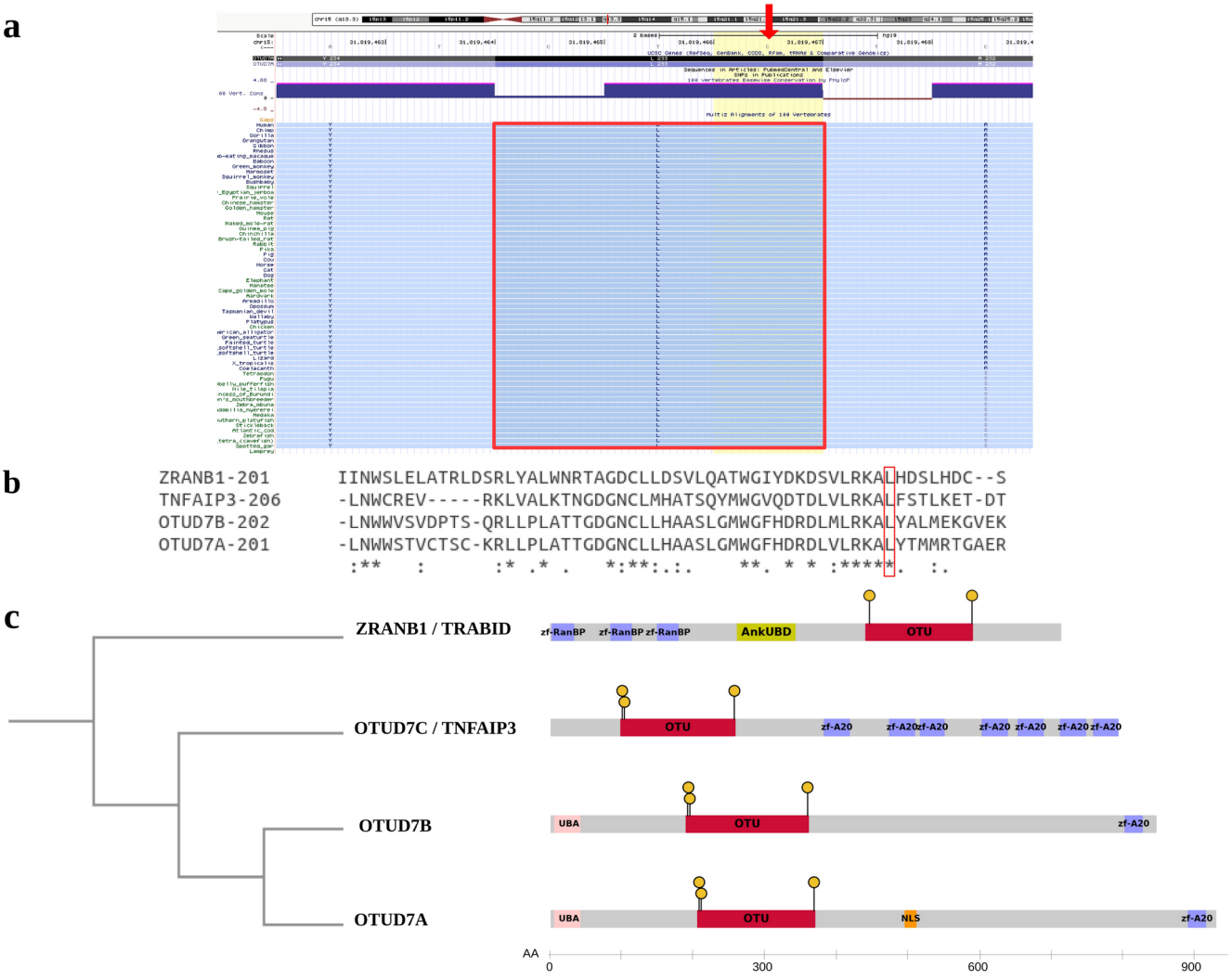


Figure 31 : Famille de protéines à domaine OTU

a) Alignement multi-espèce sur UCSC de la région mutée. La flèche rouge indique la variation identifiée chez le patient. b) Alignement par MAFFT v7 des 4 protéines de la sous-famille A20-like. Les astérisques indiquent les résidus conservés, les deux-points indiquent la conservation entre groupes de résidus à propriétés similaires et les points indiquent la conservation entre groupes de résidus aux propriétés moins proches. La position de la variation chez le patient est indiquée en rouge. c) Arbre phylogénétique selon MAFFT v7 (Katoh and Standley, 2013) et domaines des 4 protéines les plus proches de la sous-famille A20-like des déubiquitinases humaines à domaine OTU.

III.1.3- Analyse fonctionnelle de la variation

L'analyse fonctionnelle de la variation du gène *OTUD7A* a été réalisée par l'équipe de Frédéric Ebstein (Greifswald, Allemagne). La protéine OTUD7A appartient à la famille des déubiquitinases à domaine OTU. Certaines ont déjà fait l'objet d'études démontrant leur implication dans la régulation de la formation et l'activité du protéasome (Santiago-Sim et al., 2017). Nous avons donc décidé d'étudier l'impact d'un défaut de la protéine OTUD7A sur les complexes intracellulaires 20S et 26S du protéasome.

L'extrait cellulaire de fibroblastes contrôles (CTRL1) et de fibroblastes du patient, porteurs de la variation c.697C>T, a été séparé par SDS-PAGE en conditions non-dénaturantes. Les bandes correspondantes au protéasome ont été visualisées par la réaction d'hydrolyse du peptide fluorogène Suc-LLVY-AMC. La révélation du gel a révélé deux bandes à forte fluorescence correspondant aux complexes 20S et 26S/hybrides du protéasome **(Fig. 32a**). Ce gel montre également une diminution de l'activité enzymatique du complexe 20S au sein des fibroblastes du patient comparée aux fibroblastes contrôles. La mesure directe de l'activité chymotrypsin-like au sein les lysats cellulaires a donc été réalisée et confirme bien la diminution de cette activité enzymatique chez les fibroblastes mutés pour OTUD7A (**Fig. 32b**). L'expérience de Western Blot a également montré que la réduction de l'activité enzymatique n'est pas due à une diminution de la quantité de complexe 20S du protéasome (mise en évidence par la détection de la sous-unité α6) au sein des fibroblastes du patient (**Fig. 32a**). Aucune différence significative n'est observée pour la quantité de complexe 26S, mise en évidence par la détection de la sous-unité Rpn5, entre les deux lignées cellulaires. En revanche, il existe une différence significative pour la quantité de PA28 libre, régulateur du protéasome, et de complexes PA28-20S. En effet, l'anticorps dirigé contre la sous-unité PA28-α permet la mise en évidence de deux bandes aux environs de 180 kDa pour les fibroblastes contrôles, indiquant qu'une proportion importante de complexe 20S est associée avec un anneau de PA28 (**Fig. 32a**). A l'inverse, au sein des fibroblastes du patient, la révélation montre une diminution significative de PA28 libre et des complexes PA28-20S. La variation homozygote c.697C>T du patient provoquerait la baisse du nombre de complexes PA28. Afin de confirmer cette hypothèse, un Western Blot a également été réalisé sur les extraits cellulaires de 3 autres lignées

contrôles (CTRL2-CTRL4) et les fibroblastes du patient avec des anticorps dirigés contre plusieurs sous-unités du protéasome (**Fig. 32d**). Aucune variation d'expression n'est observée entre les lignées pour les sous-unités du complexe 19S (Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6, Rpn5 and Rpn10). En revanche, au sein des fibroblastes mutés pour OTUD7A, le niveau d'expression des sous-unités α2, β1, β2, β5, β1i, β2i and β5i du protéasome 20S et le PA28-α est significativement réduit. La détection des protéines ubiquitinées via la lysine 48, cible de la dégradation par le protéasome, montre que la diminution de l'activité enzymatique du protéasome est accompagnée d'une accumulation de ces protéines ubiquitinées au sein des fibroblastes du patient (**Fig. 32c**). L'ensemble de ces résultats conduit à l'hypothèse d'une relation entre les anomalies de la protéine OTUD7A et l'assemblage altéré des complexes PA28-20S par la déficience en sous-unités du protéasome 20S et de PA28.

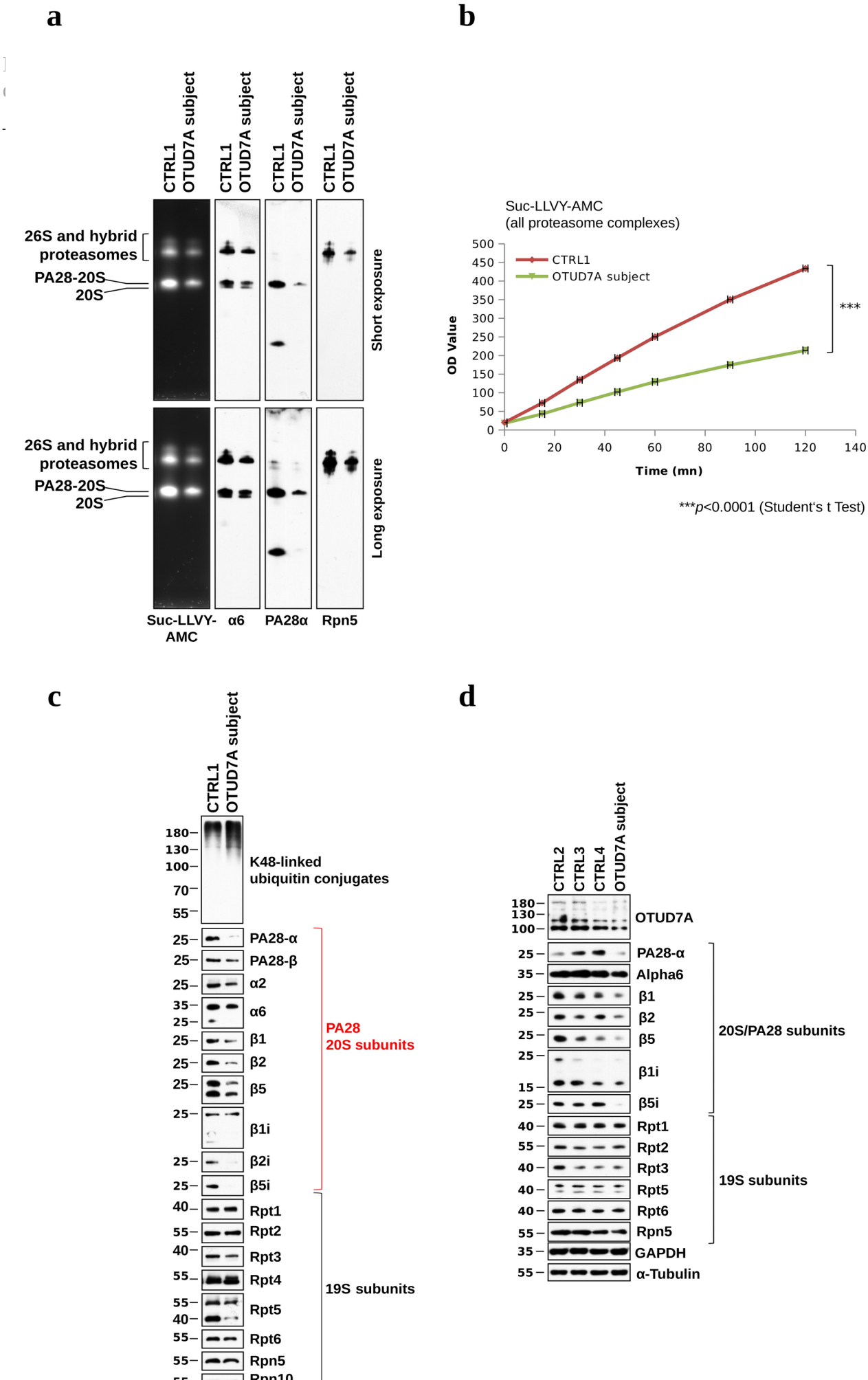

a
CTRL1
OTUD7A subject
26S and hybrid proteasomes
PA28-20S
20S
Short exposure
Long exposure
Suc-LLVY-AMC
α6
PA28α
Rpn5
b
Suc-LLVY-AMC
(all proteasome complexes)
CTRL1
OTUD7A subject
OD Value
Time (mn)
***
***p<0.0001 (Student's t Test)
c
CTRL1
OTUD7A subject
K48-linked ubiquitin conjugates
PA28-α
PA28-β
α2
α6
β1
β2
β5
β1i
β2i
β5i
PA28 20S subunits
Rpt1
Rpt2
Rpt3
Rpt4
Rpt5
Rpt6
Rpn5
Rpn10
α-Tubulin
19S subunits
d
CTRL2
CTRL3
CTRL4
OTUD7A subject
OTUD7A
PA28-α
Alpha6
β1
β2
β5
β1i
β5i
20S/PA28 subunits
Rpt1
Rpt2
Rpt3
Rpt5
Rpt6
Rpn5
19S subunits
GAPDH
α-Tubulin

Figure 32 : Impact de la variation c.697C>T du gène OTUD7A sur l'assemblage du protéasome PA28-20S dans les fibroblastes du patient

a) Western Blot des extraits cellulaires des fibroblastes contrôles (CTRL1) et du patient. Mise en évidence de l'activité chymotrypsin-like du protéasome par hydrolyse du peptide Suc-LLVY-AMC à 0,1 mM et deux temps d'exposition. Les anticorps sont dirigés contre les sous-unités α6, Rpn5 et PA28-α. b) Détermination de l'activité chymotrypsin-like par mesure de la fluorescence produite par l'hydrolyse du peptide Suc-LLVY-AMC. Les mesures ont été réalisées toutes les 15 minutes pendant 2h. Le test de Student entre les 2 lignées de fibroblastes donne une p-value inférieure à 0,0001. c) Western Blot des extraits cellulaires des fibroblastes contrôles (CTRL1) et du patient. Les anticorps sont dirigés contre les sous-unités Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6, Rpn5, Rpn10, α6, β1, β2, β5, β1i, β2i et β5i du protéasome, contre PA28-α, l'α-tubuline et les chaînes d'ubiquitines reliées par la lysine 48. d) Western Blot des extraits cellulaires de 3 autres lignées de fibroblastes contrôles (CTRL2-CTRL4) et du patient. Les anticorps sont dirigés contre OTUD7A, PA28-α et les sous-unités β1, β2, β5, β1i, β5i, Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6 et Rpn5 du protéasome. L'α-tubuline et la GAPDH sont des témoins de charge.

Afin de confirmer cette hypothèse, ces expériences ont été reproduites sur une lignée haploïde HAP1 *OTUD7A*$^{-}$ porteuse d'une délétion de 16 pb au sein de l'exon 3 du gène *OTUD7A* (NM_130901.1:c.494_509del, p.Ser165Thrfs*2). Cette lignée présente une diminution de la quantité de complexe PA28-20S comparé à la lignée HAP1 contrôle (**Fig. 33a**). Cette baisse est accompagnée d'une diminution de l'activité chymotrypsin-like du protéasome de la lignée haploïde mutée (**Fig. 33b**). Comme cela a été observé pour la lignée de fibroblastes du patient, l'expression des sous-unités du complexe 20S et de PA28-α est réduite au sein des cellules haploïdes HAP1 *OTUD7A*$^{-}$ (**Fig. 33c**). De même, il existe des agrégats de protéines ubiquitinées via la lysine 48, signe d'une anomalie dans l'activité du protéasome, au sein de la lignée haploïde déficiente en OTUD7A.

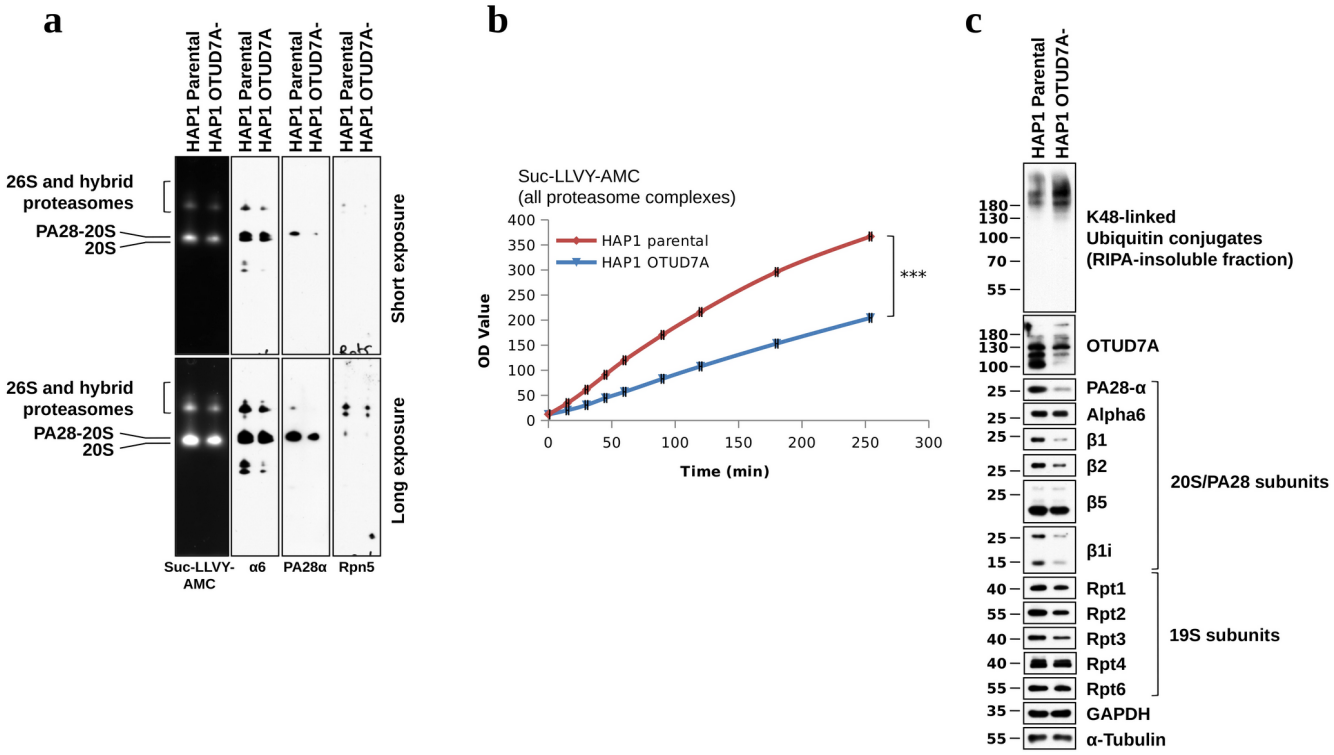


Figure 33 : Analyse fonctionnelle d'une lignée haploïde HAP1 déficiente pour OTUD7A *a) Western Blot des extraits cellulaires des lignées haploïdes contrôles (parentales) et HAP1 OTUD7A-. Mise en évidence de l'activité chymotrypsin-like du protéasome par hydrolyse du peptide Suc-LLVY-AMC à 0,1 mM et deux temps d'exposition. Les anticorps sont dirigés contre les sous-unités α6, Rpn5 et PA28-α. b) Détermination de l'activité chymotrypsin-like par mesure de la fluorescence produite par l'hydrolyse du peptide Suc-LLVY-AMC. Les mesures ont été réalisées toutes les 15 minutes pendant 4h. c) Western Blot des extraits cellulaires des lignées haploïdes contrôles (parentales) et HAP1 OTUD7A-. Les anticorps sont dirigés contre OTUD7A, PA28-α, les sous-unités Rpt1, Rpt2, Rpt3, Rpt4, Rpt6, α6, β1, β2, β5, et β1i du protéasome et les chaînes d'ubiquitines reliées par la lysine 48. L'α-tubuline et la GAPDH sont des témoins de charge.*

L'ensemble de ces travaux sur le gène *OTUD7A* a fait l'objet d'une publication dans *Clinical Genetics* (**Article 1**)

III.1.4- Discussion

Le gène *OTUD7A* est situé au sein de la région chromosomique 15q13.3. La délétion hétérozygote de cette région, qui emporte 6 gènes dont *CHRNA7* et *OTUD7A*, est à l'origine du syndrome microdélétionnel 15q13.3, de pénétrance complète et d'expressivité variable (Sharp

et al., 2008 ; Masurel-Paulet et al., 2010; Masurel-Paulet et al., 2014; Lowther et al., 2015). Le spectre des atteintes neuro-développementales de ce syndrome est large allant de la DI associée à des particularités morphologiques, de l'épilepsie, des atteintes neuropsychiatriques avec ou sans anomalies cognitives ou des difficultés d'apprentissage aspécifiques jusqu'à une absence totale de manifestations cliniques. De plus, la présentation clinique peut varier au sein d'une famille où ségrège la même délétion. Les rares descriptions de délétions homozygotes rapportent un phénotype neuro-développemental plus sévère avec encéphalopathie épileptique, hypotonie et troubles de croissance (Masurel-Paulet et al., 2010 ; Lowther et al., 2015 ; Endris et al., 2010 ; Spielmann et al., 2011 ; Simon et al., 2019 ; Uddin et al., 2018).

Bien que la délétion ait fait l'objet de plusieurs descriptions, le ou les gènes responsables du phénotype n'ont pas été clairement identifiés. Après avoir longtemps suspecté le gène *CHRNA7* d'être à l'origine du phénotype neurocognitif des patients homozygotes, hétérozygotes ou hétérozygotes composites (Soler-Alfonso et al., 2014), des études récentes sur des modèles murins ont mis en évidence le rôle potentiel du gène voisin *OTUD7A* dans le phénotype neuro-développemental du syndrome microdélétionnel (Uddin et al., 2018 ; Yin et al., 2017, 2018b). L'expression de la protéine OTUD7A dans le cerveau et sa localisation au niveau des dendrites et des compartiments synaptiques pendant la maturation des neurones corticaux renforcent cette hypothèse (Uddin et al., 2018). Cette protéine appartient à la famille des enzymes déubiquitinases (DUB) impliquées dans la régulation de nombreux processus cellulaires via le système de déubiquitination. L'ubiquitine est une protéine de 76 acides aminés présente au sein des cellules eucaryotes et qui est utilisée par ces dernières pour marquer les protéines. Cette protéine globulaire est liée à sa cible sous forme de monomère ou de polymère par les enzymes E2, E3 et E4. Cette modification post-transcriptionnelle réversible impacte l'activité, la localisation, les interactions ou le turn-over de la protéine cible (Pinto-Fernández et al., 2019). Elle permet l'adressage au protéasome de la protéine marquée en vue de sa dégradation. Au sein des cellules, la protéolyse est réalisée par deux processus : le système autophagie-lysosome et le système ubiquitine-protéasome (Rousseau and Bertolotti, 2018). Les protéines sont ciblées soit en cas d'anomalies de structure tridimensionnelle soit par ubiquitination, majoritairement par une chaîne de polyubiquitines. Il s'agit d'une chaîne d'ubiquitines reliées entre elles par une liaison entre la glycine en C-ter de chaque ubiquitine et l'une des 7 lysines de l'ubiquitine suivante (Lys6, Lys11, Lys27, Lys29, Lys33, Lys48 et Lys63)

(Komander et al., 2009). L'ubiquitination par la chaîne liée en Lys48 adresse la protéine au protéasome 26S pour dégradation (Suresh et al., 2016). Le protéasome 26S de la cellule eucaryote est un complexe composé de deux sous-parties : un cœur catalytique 20S et un complexe régulateur 19S (Rousseau and Bertolotti, 2018). Ce dernier peut être remplacé par d'autres types de complexes régulateurs, comme PA28, qui sont des heptamères. Il existe ainsi plusieurs combinaisons de complexes protéasomiques hybrides. Une fois que la protéine marquée est adressée au protéasome, elle subit une étape de déubiquitination afin de recycler l'ubiquitine nécessaire au maintien de la protéolyse (Komander et al., 2009). Ce clivage est réalisé par des enzymes DUBs. Ces dernières permettent aussi la production d'ubiquitine par clivages de ses précurseurs (Komander et al., 2009). Elles jouent également un rôle important dans la formation et le fonctionnement des synapses (Kowalski and Juo, 2012). Il existe des pathologies ayant pour origine une variation pathogène au sein d'un gène codant pour une DUB, comme par exemple, *USP9X* (OMIM 300072), *USP27X* (OMIM 300975) ou *OTUD6B* (OMIM 612021). Ces gènes sont en effet associés à la DI à transmission récessive (Homan et al., 2014; Hu et al., 2016; Santiago-Sim et al., 2017).

Trois études récentes ont permis de souligner le rôle du gène *OTUD7A* dans le phénotype du syndrome microdélétionnel. Tout d'abord Yin et al. (2017) rapportent que les souris *Chrna7*$^{-/-}$ ne présentent pas le phénotype cardinal du syndrome microdélétionnel (Yin et al., 2017). Puis l'année suivante, les mêmes auteurs étudient une lignée de souris *Otud7a*$^{-/-}$ qui deviendra le modèle murin du syndrome de la microdeletion 15q13.3, renforçant ainsi l'hypothèse de l'implication d'*OTUD7A* dans la pathologie. La même année, Uddin et al. montrent qu'il s'agit d'un gène majeur impliqué dans le syndrome. Parallèlement à l'étude d'une lignée de souris Df(h15q13)/+, les auteurs ont rapporté 2 nouveaux patients atteints de TSA et 1 nouveau patient atteint de retard global du développement, porteurs de variations au sein d'*OTUD7A*. Les deux premiers sont porteurs d'une délétion *de novo* en phase de 9 pb (p.Asn492_Lys494del). Le troisième possède une microdélétion de la région 15q13.3 située entre les points de cassure BP4 et BP5 incluant *OTUD7A* mais pas *CHRNA7* ni sa région en amont. Après avoir réétudié des cas de délétions du gène *CHRNA7* sans *OTUD7A*, les auteurs suggèrent que le phénotype des patients concernés serait dû à une anomalie dans l'expression du gène *OTUD7A* plutôt qu'à un dysfonctionnement de sa protéine. En effet, *CHRNA7* et *OTUD7A* sont en anti-sens et partagent la même région en amont de leurs promoteurs. Cette

région possède un site de fixation pour EZH2 un facteur de transcription. Ce dernier est impliqué dans la croissance dendritique et régulerait donc l'expression du gène *OTUD7A* (Qi et al., 2014). Les délétions du seul gène *CHRNA7* emportent également cette région en amont, ce qui a conduit à suspecter initialement ce gène comme responsable du phénotype observé.

Le patient identifié au cours de ces travaux comme porteur de la variation homozygote NM_130901.2:c.697C>T, p.Leu233Phe dans le gène *OTUD7A* est le premier cas rapporté de ce type. Sa présentation clinique est en adéquation avec celles des patients rapportés comme ayant une microdélétion 15q13.3 homozygote ou hétérozygote composite : retard global et sévère du développement, retard de langage et spasmes infantiles se transformant en crises d'absence atypiques.

Aucune variation additionnelle n'a été identifiée comme ayant un impact sur le phénotype du patient. Il a été montré que des variations au sein de gènes codant pour des DUBs peuvent être à l'origine de pathologies neuro-développementales. Leurs effets sur la protéolyse ont donc mené à une collaboration avec l'équipe du Dr Frédéric Ebstein (Greifswald, Allemagne) afin de réaliser des tests fonctionnels pour identifier l'impact de la variation homozygote sur ces mécanismes moléculaires. Les résultats ont permis de mettre en évidence un lien entre le gène *OTUD7A* et le fonctionnement du protéasome. Ainsi, la protéine OTUD7A contrôlerait l'expression de PA28, régulateur du protéasome, et des sous-unités α et β.

Les données obtenues à partir des fibroblastes du patient, porteurs de la variation homozygote c.697C>T du gène *OTUD7A*, mettent en évidence une diminution de la quantité de PA28 et de sous-unités α et β du protéasome par rapport aux lignées contrôles (**Figs. 32c et 32d**). Cette diminution est accompagnée d'une réduction de la formation de complexes protéasomiques PA28-20S (**Fig. 32a**). Ces derniers sont impliqués dans la dégradation des protéines oxydées (Pickering et al., 2010). Un défaut de fonctionnement du protéasome pourrait conduire à l'accumulation d'agrégats de protéines insolubles à l'origine d'un stress oxydatif pour la cellule. Il a été montré que le stress oxydatif contribue à l'apparition de pathologies neuro-développementales ou neuro-dégénératives (Reeg and Grune, 2015 ; Adav and Sze, 2016 ; Wells et al., 2016).

L'étude d'une lignée de cellules haploïdes HAP1 *knockout* pour *OTUD7A* a mis en

évidence une diminution de l'assemblage des complexes protéasomiques PA28-20S et une augmentation de l'accumulation de protéines ubiquitinées (**Fig. 33c**). Cette lignée haploïde reproduit donc le phénotype des fibroblastes du patient. Ces observations permettent de démontrer l'existence d'une relation de cause à effet entre la déficience en protéine OTUD7A et les anomalies d'assemblage et de fonctionnement du protéasome. Néanmoins, le ou les mécanismes par lesquels OTUD7A influence l'expression de PA28 et des sous-unités α et β restent méconnus. Des travaux supplémentaires seront nécessaires pour déterminer si la réduction, au sein des cellules déficientes pour OTUD7A, de la quantité de PA28 et des sous-unités α et β est due à la diminution de la transcription et/ou de la traduction ou à l'augmentation du turn-over protéique. La détermination de la quantité d'ARNm de ces protéines par qPCR pourrait permettre de mettre en évidence une diminution de la transcription. Si aucun défaut de la transcription n'était observé mais qu'il existait une diminution de la quantité de protéine (observée par Western Blot) il s'agirait d'une anomalie de la traduction ou de la stabilité des protéines étudiées. Une étude a été réalisée en 2017 par Santiago-Sim et al. sur les cellules d'un patient avec atteintes neurocognitives et porteur d'une délétion biallélique au sein d'un autre gène de la famille OTU, le gène *OTUD6B* (Santiago-Sim et al., 2017). Ces cellules sont incapables d'incorporer les sous-unités 19S au sein des complexes 26S du protéasome. Les résultats obtenus lors de cette thèse renforcent donc l'hypothèse selon laquelle les déubiquitinases à domaine OTU participent activement à la régulation de l'assemblage et/ou de la quantité des isoformes du protéasome (Santiago-Sim et al., 2017). Mais contrairement à *OTUD6B*, le gène *OTUD7A* semble nécessaire à l'expression de PA28. La protéine OTUD7A clive spécifiquement les chaînes d'ubiquitine liées en Lys11 (Mevissen et al., 2013, 2016). Une des hypothèses serait donc que l'expression de PA28 nécessiterait le clivage de ce type de chaînes liées à PA28 et/ou à d'autres régulateurs du protéasome. Des analyses supplémentaires permettraient de décrire le profil d'ubiquitination de PA28 et son impact sur sa quantité.

Ces résultats d'analyse fonctionnelle ont permis mettre en évidence le rôle du gène *OTUD7A* dans les atteintes neurocognitives. Le patient étudié porteur d'une variation homozygote dans ce gène présente un phénotype similaire à ceux des patients homozygotes ou hétérozygotes composites pour la microdélétion 15q13.3. Le travail de lecture en recherche de cet exome, associé à ces études fonctionnelles, permet de considérer le gène *OTUD7A* comme

un candidat solide pour expliquer le phénotype de la microdélétion 15q13.3 et celui de l'encéphalopathie épileptique précoce décrite chez le patient.

Parallèlement à ces travaux de recherche sur le gène *OTUD7A,* le travail de lecture en recherche a identifié, chez un second patient, un autre gène candidat pour la déficience intellectuelle avec syndrome autistique : le gène *DLGAP2*.

III.2- Identification d'une variation hétérozygote non-sens au sein du gène *DLGAP2*

III.2.1- Présentation clinique de la patiente

La patiente est née à 41 SA de parents originaires d'Afrique du Nord et non apparentés. Aucun antécédent familial n'a été décrit, il s'agit d'un cas sporadique. Ses mensurations à la naissance étaient dans la moyenne. La marche a été acquise à 12 mois. Lors de la dernière consultation à 6 ans et demi, ses mensurations étaient également dans la moyenne. La patiente a été adressée pour déficience intellectuelle avec syndrome autistique associé à un retard du développement, des troubles de la motricité fine et du langage, un trouble de déficit de l'attention avec hyperactivité et un comportement agressif. L'IRM cérébrale était normale. Une analyse en CGH-array et un caryotype ont été réalisés en amont de l'analyse de l'exome, sans mettre en évidence d'anomalie. De même, la recherche des syndromes de l'X fragile et d'Angelman étaient négatifs. Enfin, un panel de gènes comprenant les gènes *CDKL5*, *IQSEC2*, *MBD5*, *MEF2C*, *SLC9A6*, *STXBP1*, *TCF4*, *UBE3A*, *ZEB2* (encéphalopathies épileptiques) et *MECP2* n'a pas mis en évidence de variation pathogène.

III.2.2- Réanalyse recherche de l'exome

L'analyse dans un cadre de recherche de l'exome de la patiente a mis en évidence une variation hétérozygote tronquante (confirmée *de novo* par séquençage Sanger) dans le gène *DLGAP2* (OMIM 605438) : chr8:g.1624744G>T (NM_004745.4:c.2008G>T,p.(Glu670*)) (**Fig. 34a**). Cette variation était absente de la base de données gnomAD et les scores bioinformatiques de prédictions étaient en faveur d'une pathogénicité (CADD=39, GERP=5.66, PolyPhen2 non applicable, MISZ=-0,895 et pLI=0.950). De plus, la variation tronquante est à l'origine de la perte du domaine GKAP (Guanylate-kinase-associated protein) conservé au sein de la famille DLGAP et de nombreuses espèces animales (Rasmussen et al., 2017) (**Fig 34b**). De plus, les scores de prédiction fonctionnelle SIFT et Provean (sur Varsome (Kopanos et al., 2019)) prédisaient un impact négatif de la variation. Enfin, la protéine DLGAP2 interagit avec d'autres protéines produites par des gènes déjà impliqués en pathologies humaines de type DI ou autisme (Rasmussen et al., 2017) (**Fig. 34c**).

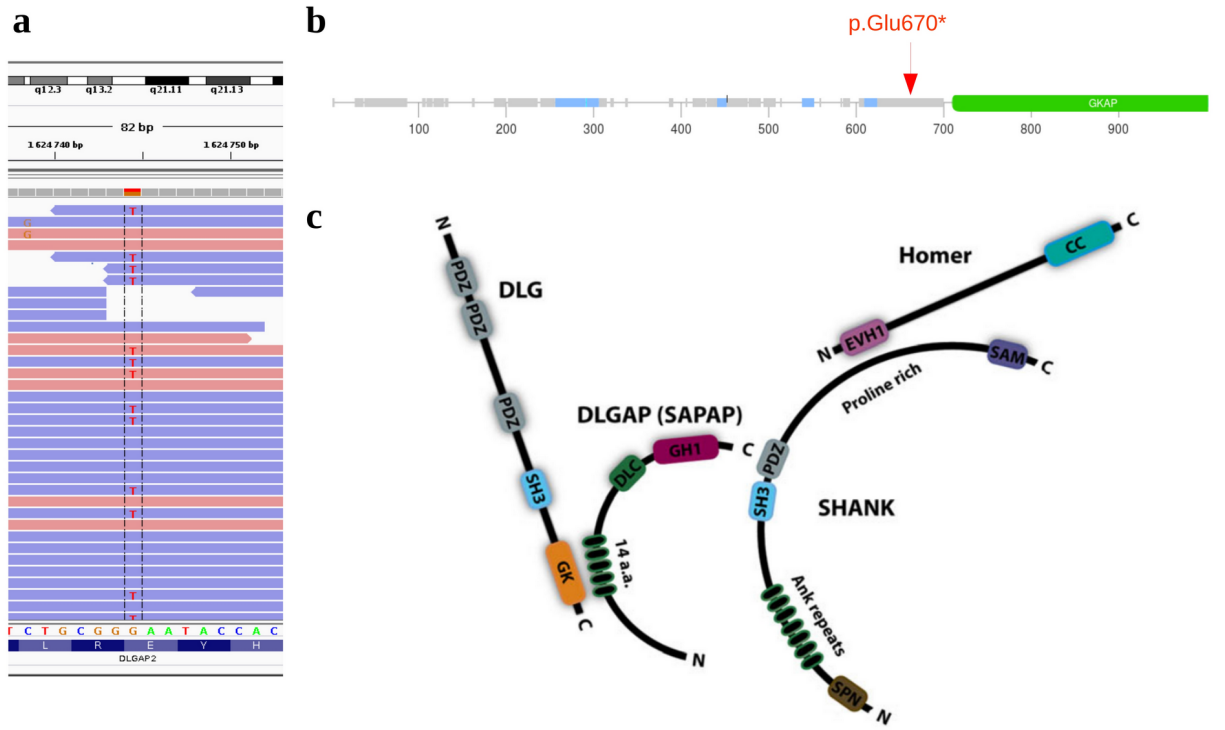


Figure 34 : Variation hétérozygote au sein du gène DLGAP2
a) Vue IGV de la variation chr8:g.1624744G>T (NM_004745.4:c.2008G>T, p.(Glu670)) identifiée par approche d'exome solo chez la patiente. b) Localisation de la variation 39 acides aminés en N-terminal du domaine GKAP (Guanylate-Kinase-Associated Protein) conservé au sein de la famille DLGAP et de nombreuses espèces animales (Rasmussen et al., 2017). c) Interaction via l'extrémité C-terminale des protéines DLGAPs avec les protéines SHANKS. Interactions des DLGAPs avec des produits des gènes DLG3 impliqués dans la DI liée à l'X (OMIM 300189), SHANK2 impliqué dans une susceptibilité à l'autisme (OMIM 603290) et SHANK3 impliqué dans le syndrome Phelan-McDermid ou la schizophrénie (OMIM 606230) (d'après Rasmussen et al., 2017).*

III.2.3- Collaboration internationale

Ce gène candidat a donc été partagé sur GeneMatcher, site de partage de variations candidates dans le but de constituer des cohortes de patients de même pathologie (Sobreira et al., 2015) Elle a permis de nous mettre en relation avec plusieurs équipes ayant identifié des variations dans ce gène chez des patients avec un retard neurodéveloppemental. Les variations identifiées étaient des variations faux-sens bi-alléliques ou des variations tronquantes *de novo*. L'équipe de Davor Lessel (Hambourg) est en cours de collecte de patients porteurs de

variation(s) dans le gène *DLGAP2* en vue d'une publication collaborative, à laquelle notre équipe sera associée.

III.2.4- Discussion

Le gène *DLGAP2* (OMIM : 605438) est localisé dans la région 8p23.3 décrite comme étant impliquée dans des pathologies de type schizophrénie (The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium et al., 2011). Plusieurs patients, atteints de Trouble du Spectre Autistique (TSA), ont été décrits comme porteurs d'une duplication ou d'une délétion de cette région génomique et du gène *DLGAP2*. En effet, la duplication de 791kb décrite par Marshall et al. (Marshall et al., 2008) met en évidence l'implication de la région 8p23.3 dans le TSA. Le gène *DLGAP2* est alors suspecté d'être impliqué dans le phénotype du patient atteint de retard du langage, de comportements répétitifs et de TSA. En 2010, Chien et son équipe décrivent un patient atteint d'autisme, d'hyperactivité, d'inattention, d'impulsivité en lien avec un syndrome autistique (Chien et al., 2010). Ce garçon présente une délétion 8p23.2-pter de 2,4Mb. Parmi les gènes pouvant être associés à ce phénotype, Chien et al., retiennent *DLGAP2*. En effet les auteurs soulignent la présence de ce gène dans la région dupliquée chez un autre patient autistique (Marshall et al., 2008) et délétée chez une femme également atteinte de TSA (Ozgen et al., 2009).

Pour renforcer cette hypothèse, Poquet et al. (Poquet et al., 2017) comparent les duplications du bras court du chromosome 8 chez 6 nouveaux patients ainsi que les patients décrits par Pinto et al. (2010) et Marshall et al (2008). Ils mettent en évidence que la région commune de ces CNV est le locus de *DLGAP2*. Parmi les 6 nouvelles délétions décrites une est *de novo*, 3 sont héritées du père non atteint, une est héritée de la mère apparemment non atteinte et une est héritée de la mère présentant une atteinte modérée.

Par ailleurs, des variations ponctuelles au sein du gène *DLGAP2* ont également été rapportées. Chien et al. identifient des variations prédites comme pathogènes chez des patients atteints de TSA (Chien et al., 2013). Bien qu'ils suggèrent une cause multifactorielle à l'apparition du phénotype autistique chez leurs patients, ils conservent l'hypothèse selon

laquelle *DLGAP2* est impliqué dans l'autisme. Puis Iossifov et al., identifient une variation candidate *de novo* chez un patient atteint de TSA (Iossifov et al., 2014). A leur tour, Li et al. mettent en évidence, chez des patients atteints de schizophrénie, des variations prédites pathogènes dans le gène *DLGAP2* ainsi que des variations augmentant l'activité du promoteur du gène (Li et al., 2014). Ils suggèrent alors que l'augmentation du taux de protéine DLGAP2 chez ces derniers conduit au phénotype schizophrénique.

Outre les CNV et SNV impliquant le gène *DLGAP2*, l'étude du modèle de souris *knockout* pour *Dlgap2* renforce l'hypothèse de l'implication de *DLGAP2* dans des phénotypes neurologiques allant du TSA aux troubles de la coordination (Jiang-Xie et al., 2014). En effet l'étude détaillée du phénotype a mis en évidence un comportement de type autistique chez cette lignée de souris avec agressivité, hyperactivité et dysfonctionnement social. De plus l'analyse moléculaire a montré que ces souris *Dlgap2*$^{-/-}$ présentaient une réduction significative du nombre de récepteurs synaptiques AMPA (GluR1) et NMDA (NR1) du cortex cérébral. Il a également été mis en évidence une diminution du niveau de protéines de structure Homer-1b/c et αCaMKII pouvant être à l'origine d'un défaut de structure synaptique, dont l'intégrité est nécessaire au bon fonctionnement de cette dernière. Par ailleurs, ils ont étudié la densité des épines des neurones pyramidaux du cortex orbitofrontal (COF) car il a été montré qu'un dysfonctionnement de ce dernier, chez l'humain, les primates ou les rongeurs, amplifie le comportement agressif. Ainsi les mesures effectuées sur ces souris *Dlgap2*$^{-/-}$ ont mis en évidence une réduction de la densité des épines neuronales à l'origine d'une diminution du signal excitateur du COF reçu par les neurones pyramidaux**.** L'analyse par microscopie électronique a également montré une diminution significative de la taille et de l'épaisseur de la densité post-synaptique dans le COF. La structure post-synaptique au sein du COF est donc déficitaire chez les souris *Dlgap2*$^{-/-}$, gène régulant la synaptogenèse au sein du COF en tant qu'un des principaux composants des épines dendritiques. Parallèlement à l'analyse de structure, Jiang-Xie et al. ont étudié le fonctionnement des synapses de cette lignée de souris (Jiang-Xie et al., 2014). Les expériences d'électrophysiologie ont mis en évidence des réponses post-synaptiques plus faibles au niveau des synapses existantes. Mais il a également été observé une diminution de la libération de vésicules présynaptiques. Les auteurs suggèrent que cette observation soit liée à une signalisation rétrograde ou via des molécules d'adhésion trans-synaptiques ; car DLGAP2 est une protéine post-synaptique. L'étude moléculaire et

phénotypique des souris *Dlgap2*$^{-/-}$ souligne le rôle important de DLGAP2 dans le maintien et la structure de la densité post synaptique au niveau du COF, région cérébrale impliquée dans l'inhibition comportementale et la régulation des tendances agressives**.**

Plus récemment, Rasmussen et al. ont publié une étude sur l'expression des gènes *DLGAPs* et le rôle de leurs protéines dans le fonctionnement cérébral en particulier au niveau des synapses, en lien avec des pathologies cérébrales (Rasmussen et al., 2017). Ils ont mis en évidence les anomalies de la densité post-synaptique chez les souris *Dlgap2*$^{-/-}$. Ces études renforcent l'hypothèse de l'implication du gène *DLGAP2* dans le TSA.

DLGAP2 est une protéine exprimée au niveau des testicules, de la glande pituitaire et majoritairement au niveau du cerveau, plus particulièrement dans le cortex, l'hippocampe, le bulbe olfactif et le striatum (Rasmussen et al., 2017). Elle comporte 3 domaines (**Fig. 34c**) :

- un domaine répété de 14 acides-aminés, permettant l'interaction avec le domaine GK des protéines DLG1, DLG2 et DLG4. Ces protéines sont d'ailleurs suspectées d'être impliquées dans la schizophrénie et le TSA (Xing et al., 2016)**.**
- un domaine DLC interagissant directement avec la protéine motrice DLC présente au niveau des dendrites et densité post-synaptique (Naisbitt et al., 2000)**.**
- un domaine GH1 permettant l'interaction avec les protéines de la famille SHANKs. (**Fig. 34**) conduisant ainsi à l'intégrité du complexe protéique DLG-DLGAPs-SHANK nécessaire au maintien de la structure et au bon fonctionnement des synapses excitatrices (Kim et al., 1997)**.** Les protéines SHANKs sont importantes pour la maturation de la densité post-synaptique. Elles sont d'ailleurs décrites comme impliquées dans la schizophrénie (SHANK3) et le TSA (SHANK1-3) (Guilmatre et al., 2014)**.**

De plus, la structure des protéines DLGAPs est bien conservée entre les espèces et au sein de cette famille protéique (**Fig. 34c**). Leur fonction est donc très probablement essentielle et nécessite l'intégrité des protéines. Enfin, elles sont toutes suspectées d'être associées à des pathologies neurologiques de type schizophrénie, autisme ou bipolarité.

Pour renforcer l'hypothèse d'une association entre *DLGAP2* et le phénotype de la patiente concernée, un partage sur GeneMatcher a été réalisé. Il a permis d'identifier des

patients chez lesquels une variation du gène *DLGAP2* est également suspectée d'être à l'origine du retard neuro-développemental. Les variations *de novo* identifiées sont toutes des variations perte de fonction tandis que les faux-sens sont tous bialléliques. Ainsi la présentation clinique de notre patiente et la ségrégation de la variation sont en adéquation avec celles des autres patients identifiés. La collaboration internationale est gérée par le Dr Davor Lessel (Hambourg) qui travaille en collaboration avec l'équipe du Pr Stephan Kindler (Hambourg). Ce dernier étudie les protéines de la famille DLGAPs. Son équipe a travaillé sur le modèle de souris *knockout* pour *Dlgap4* (Schob et al., 2019) et a récemment généré un modèle de souris *knockout* pour *Dlgap2* en cours de caractérisation. Cette collaboration va permettre d'apporter des arguments supplémentaires quant au rôle de *DLGAP2* dans un phénotype neuro-développemental. Une publication est en cours de rédaction.

IV- DISCUSSION GÉNÉRALE : LA RÉANALYSE RECHERCHE, UNE STRATÉGIE POUR IDENTIFIER DE NOUVEAUX GÈNES IMPLIQUÉS DANS LES AD

Le travail de lecture, dans un cadre de recherche, d'exomes négatifs en diagnostic a permis l'identification de deux gènes candidats parmi 80 exomes réanalysés au sein de la cohorte de patients séquencés en exome diagnostic par le laboratoire CERBA, *OTUD7A* et *DLGAP2*, soupçonnés d'être respectivement impliqués dans l'encéphalopathie épileptique et la déficience intellectuelle avec syndrome autistique. Un troisième diagnostic a également été réalisé avec une variation dans le gène *RAC1* devenu OMIM morbide entre les 2 analyses. Le fichier VCF n'ayant pas été réannoté avec les nouvelles données d'OMIM, c'est l'approche en recherche qui a permis de réaliser le diagnostic. Cette approche de recherche est toujours précédée d'une réanalyse diagnostique. Elle a permis ici d'identifier une variation supplémentaire (chez 1 patient). Il s'agit d'une variation dans le gène *DDX3X*, déjà connu en pathologie humaine, qui n'avait pas été considérée lors de la première lecture d'exome. Après

confirmation Sanger, le diagnostic a été rendu à la famille. Il y a donc eu une amélioration de 5 % du taux diagnostique au sein de cette cohorte.

Avec environ 7-8 % de diagnostics supplémentaires, la réanalyse des gènes connus en pathologie humaine a montré son intérêt dans la réduction de l'errance diagnostique. Wenger et al. (2017) ou Nambot et al. (2018) rappellent ainsi l'importance d'une réanalyse régulière des exomes négatifs en réannotant les variations avec les mises à jour des bases de données. La lecture des exomes négatifs du Laboratoire Cerba a toujours débuté par une relecture des variations présentes au sein des gènes OMIM morbides. Certains fichiers n'ont pas été réannotés d'où la découverte d'une variation dans un gène nouvellement connu (*RAC1*) par l'approche recherche et le faible pourcentage d'amélioration.

La réanalyse des gènes OMIM morbides permet également de repérer des variations non retenues car mal interprétées ou filtrées. Al-Nabhani et al. (2018) ont ainsi augmenté leur taux diagnostique de 2 % (1 patient) et Nambot et al. (2018) de 2,5 % (4 patients). Avec une variation reclassée (*DDX3X*) au sein de la cohorte du Laboratoire Cerba le taux diagnostique a augmenté de 1,25 %.

Toutes les équipes effectuant une réanalyse des gènes connus en pathologies humaines ne se tournent pas vers une approche recherche (Wenger et al., 2017). Pourtant elle permet d'améliorer les connaissances et de repousser cette limite clinico-biologique en renforçant l'implication de nouveaux gènes en pathologie humaine. L'étude menée par l'équipe GAD de juillet 2013 à décembre 2017 a montré que cette approche permet d'augmenter le taux diagnostique d'environ 7% et a conduit à l'identification de 17 nouveaux gènes (Bruel et al., 2019a) (**Fig. 35**). Une autre étude de l'équipe présente le même taux de diagnostics supplémentaires (7,7 %) réalisés par une approche de recherche (Nambot et al., 2018). Pour Al-Nabhani et al. (2018) la réanalyse recherche n'a augmenté que de 2 % leur taux diagnostique. Avec 2 nouveaux diagnostics pour 80 patients de la cohorte du laboratoire Cerba, l'amélioration a été de 2,5 %. Il est à noter que ces 2 cohortes ne disposaient pas ou partiellement d'un pipeline de détection des CNV. Dans les 2 études de l'équipe il a été montré que ce dernier permettait d'augmenter de 1 à 2 % le taux diagnostique. D'autres analyses annexes (ex : disomies, variations en mosaïque,etc.) ont également été réalisées par certaines équipes comme

celle de Wright qui a amélioré de 8 % le nombre de diagnostics (Wright et al., 2018). Ce type d'analyse nécessite des outils supplémentaires qui ne sont pas déployés au sein de toutes les équipes.

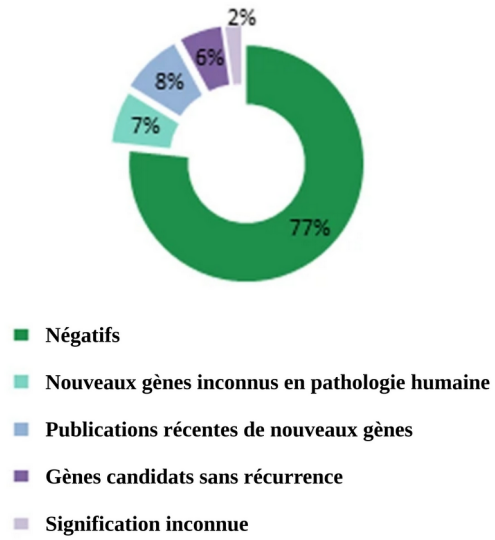


Figure 35 : Augmentation du taux diagnostique par réanalyse recherche d'exomes négatifs
La réanalyse recherche a permis d'impliquer de nouveaux gènes non connus en pathologie humaine dans 7 % de la cohorte. Des gènes et/ou phénotypes nouvellement publiés mais non référencés dans la base de données OMIM ont été détectés chez 8 % des individus. Les gènes candidats sans autre cas identifiés en interne ou à l'internationale (sans récurrence) ont été mis en évidence dans 6 % de la cohorte. Pour 2 % de la cohorte, la ségrégation n'a pas pu être réalisée rendant difficile l'implication des nouveaux gènes candidats dans une pathologie humaine (d'après (Bruel et al., 2019a)).

Le partage de données constitue une approche intéressante à développer notamment parce qu'il s'agit de regrouper des patients porteurs d'une variation génétique dans le même gène. Cette approche a été utilisée par Eldomery et al. (2017) et Nambot et al. (2018) pour faciliter l'implication de nouveaux gènes en pathologie humaine. L'équipe GAD participe activement à ce partage via GeneMatcher qui a permis l'identification, en 30 mois, de 23 nouveaux gènes impliqués dans des AD et/ou DI rares (Bruel et al., 2019b).

Mais plus de la moitié des exomes relus en diagnostic ou dans un cadre de recherche restent négatifs. Repousser les limites bioinformatiques permettrait d'augmenter le nombre de résultats positifs. En effet, comme rappelé plus tôt, il a été montré que des analyses annexes peuvent conduire à l'augmentation du taux de résolution des exomes (Wright et al., 2018). Il

s'agit d'étudier des anomalies, autres que les SNV voire CNV que les pipelines dits « classiques » détectent déjà. La détection et l'analyse de ces anomalies supplémentaires nécessitent néanmoins l'amélioration des pipelines bioinformatiques notamment par le développement d'outils spécifiques et l'ajout de fonctionnalités additionelles d'analyse de données de séquençage à haut débit, thématique qui a été l'objet de la suite de cette thèse.

Article 1

Report of the first patient with a homozygous *OTUD7A* variant responsible for epileptic encephalopathy and related proteasome dysfunction.

Clin Genet, Apr 2020;07(4),567-575

Philippine Garret, Frédéric Ebstein, Geoffroy Delplancq, Blandine Dozieres-Puyravel, Aïcha Boughalem, Stéphane Auvin, Yannis Duffourd, Sandro Klafack, Barbara A. Zieba, Sana Mahmoudi, Karun K. Singh, Laurence Duplomb, Christel Thauvin-Robinet, Jean-Marc Costa, Elke Krüger, Detlef Trost, Alain Verloes, Laurence Faivre, Antonio Vitobello.

Received: 17 September 2019 | Revised: 18 January 2020 | Accepted: 20 January 2020

DOI: 10.1111/cge.13709

CLINICAL GENETICS WILEY

ORIGINAL ARTICLE

Report of the first patient with a homozygous *OTUD7A* variant responsible for epileptic encephalopathy and related proteasome dysfunction

Philippine Garret[1,2] | Frédéric Ebstein[3] | Geoffroy Delplancq[1,4] | Blandine Dozieres-Puyravel[5] | Aïcha Boughalem[2] | Stéphane Auvin[5,6] | Yannis Duffourd[1,4] | Sandro Klafack[3] | Barbara A. Zieba[3] | Sana Mahmoudi[7] | Karun K. Singh[8] | Laurence Duplomb[1,4] | Christel Thauvin-Robinet[1,4,9] | Jean-Marc Costa[2] | Elke Krüger[3] | Detlef Trost[2] | Alain Verloes[6,10] | Laurence Faivre[1,11] | Antonio Vitobello[1,4]

[1]UMR1231 GAD, Inserm – Université Bourgogne-Franche Comté, Dijon, France

[2]Laboratoire CERBA, Saint-Ouen l'Aumône, France

[3]Universitätsmedizin Greifswald, Institut für Medizinische Biochemie und Molekularbiologie, Greifswald, Germany

[4]Unité Fonctionnelle Innovation en Diagnostic génomique des maladies rares, FHU-TRANSLAD, CHU Dijon Bourgogne, Dijon, France

[5]AP-HP, Hôpital Robert Debré, Service de Neurologie pédiatrique, Paris, France

[6]UMR1141 INSERM, Université Paris Diderot, Paris, France

[7]Service de Pédiatrie, Centre Hospitalier René-Dubos, Pontoise, France

[8]Department of Biochemistry and Biomedical Sciences, Stem Cell and Cancer Research Institute, McMaster University, Hamilton, Canada

[9]Centre de Référence Maladies Rares "déficience intellectuelle", centre de génétique, FHU-TRANSLAD, CHU Dijon Bourgogne, Dijon, France

[10]Genetics Department, AP-HP, Robert-Debré University Hospital, Paris, France

[11]Centre de Référence Maladies Rares "Anomalies du développement et syndromes malformatifs", centre de génétique, FHU-TRANSLAD, CHU Dijon Bourgogne, Dijon, France

Correspondence
Laurence Faivre and Antonio Vitobello, UMR1231 GAD, Inserm – Université Bourgogne-Franche Comté, Dijon, France.
Email: laurence.faivre@chu-dijon.fr (L. F.), antonio.vitobello@u-bourgogne.fr (A. V.)

Funding information
Association Nationale de la Recherche et de la Technologie; Conseil régional de Bourgogne-Franche-Comté; FEDER 2017; Fritz-Thyssen Foundation, Grant/Award Number: Az: 10.16.2.022MN; German Research Foundation, Grant/Award Number: SFBTR 167; Molecular Medicine Research Consortium of the University of Greifswald, Grant/Award Number: FOVB-2018-11; PARI 2017

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1111/cge.13709/.

Abstract

Heterozygous microdeletions of chromosome 15q13.3 (*MIM: 612001*) show incomplete penetrance and are associated with a highly variable phenotype that may include intellectual disability, epilepsy, facial dysmorphism and digit anomalies. Rare patients carrying homozygous deletions show more severe phenotypes including epileptic encephalopathy, hypotonia and poor growth. For years, *CHRNA7* (MIM: 118511), was considered the candidate gene that could account for this syndrome. However, recent studies in mouse models have shown that *OTUD7A/CEZANNE2* (*MIM: 612024*), which encodes for an ovarian tumor (OTU) deubiquitinase, should be considered the critical gene responsible for brain dysfunction. In this study, a patient presenting with severe global developmental delay, language impairment and epileptic encephalopathy was referred to our genetics center. Trio exome sequencing (tES) analysis identified a homozygous *OTUD7A* missense variant (NM_130901.2: c.697C>T), predicted to alter an ultraconserved amino acid, p.(Leu233Phe), lying within the OTU catalytic domain. Its subsequent segregation analysis revealed that

Philippine Garret, Frédéric Ebstein, Geoffroy Delplancq, and Blandine Dozieres-Puyravel contributed equally to this work.

the parents, presenting with learning disability, and brother were heterozygous carriers. Biochemical assays demonstrated that proteasome complex formation and function were significantly reduced in patient-derived fibroblasts and in *OTUD7A* knockout HAP1 cell line. We provide evidence that biallelic pathogenic *OTUD7A* variation is linked to early-onset epileptic encephalopathy and proteasome dysfunction.

KEYWORDS

15q13.3 microdeletion, CHRNA7, OTUD7A, proteasome

1 | INTRODUCTION

Typical chromosome 15q13.3 deletion is caused by nonallelic homologous recombination between low copy repeat elements that cluster into six breakpoint regions (BP1-BP6). Deletions mediated by breakpoint 4 (BP4) and breakpoint 5 (BP5) result in 15q13.3 microdeletion syndrome (OMIM 612001; Figure S1), leading to the loss of a 1.5 Mb region containing the critical genes *OTUD7A* (MIM: 612024) and *CHRNA7* (MIM: 118511). Chromosome 15q13.3 microdeletion syndrome is associated with a wide range of neurodevelopmental disorders.[2,3] These range from mental retardation with dysmorphic features, epilepsy, neuropsychiatric disturbances with or without cognitive impairment to unspecific learning difficulties or to the complete absence of anomalies, even within families segregating the same deletion. Visual impairment and retinal dysfunction have been observed in cases where the *TRPM1* gene is also included in the deletion.[4,5] Although 15q13.3 microdeletions show highly variable expressivity and incomplete penetrance, almost 80% of carriers manifest a neurodevelopmental/neuropsychiatric condition including developmental delay (DD)/intellectual disability (ID), epilepsy, speech problems, autism spectrum disorder (ASD), schizophrenia, mood disorders, and attention deficit hyperactivity disorder.[5] Homozygous or compound heterozygous (ie, the condition implicating the presence of two different mutated alleles at a particular gene locus) 15q13.3 microdeletions have occasionally been reported.[1,3,6] These individuals present more severe neurodevelopmental problems, with epileptic encephalopathy, hypotonia and poor growth.[1,3,5-8] Although the classical microdeletion had been widely reported in a heterozygous state, the genes responsible for the clinical features associated with 15q13.3 microdeletion syndrome have not clearly been identified.

Mouse models for 15q13.3 microdeletion syndrome replicate many of the cardinal features observed in humans.[9,10] For a long time, *CHRNA7* was considered the main candidate gene accounting for the neuropsychiatric and behavioral disorders associated with 15q13.3 copy number variation.[11] Encoding the alpha-7 neuronal nicotinic receptor subunit, CHRNA7 is a member of a superfamily of ligand-gated ion channels mediating signal transduction, synaptic plasticity, learning and memory.[12,13] Patients carrying compound heterozygous or homozygous microdeletions encompassing only the *CHRNA7*[14,15] gene are reported to have a neurodevelopmental phenotype similar to those associated with the larger 15q13.3 microdeletion.[14] However, recent studies from Yin et al.[16] showed that *Chrna7 knockout* (*Chrna7*$^{-/-}$) mice manifest no consistent neuropsychiatric or behavioral phenotypes, whereas *Otud7a knockout* (*Otud7a*$^{-/-}$) mice exhibit many of the neurological features of 15q13.3 microdeletion syndrome.[17] *OTUD7A/CEZANNE2* encodes for an ovarian tumor (OTU) deubiquitinase, which specifically cleaves Lys11-linked ubiquitin chains.[18] Uddin et al.[8] showed that OTUD7A localizes to the dendritic spines of cortical neurons, and its reduced levels in 15q13 microdeletion contributes to the dendritic spine and dendrite outgrowth deficits observed in mouse models with a syntenic heterozygous deletion (Df[h15q13]/+), supporting the idea that *OTUD7A* is a major regulatory gene for 15q13.3 microdeletion syndrome. Interestingly, biallelic variants of *OTUD6B*, another member of the OTU family, were recently reported as being responsible for a syndrome associating ID, seizures and dysmorphic features[19] via a 26S proteasome assembly defect, indicating that multiple members of this family may be involved in neurodevelopment.

In support of the findings described in mice, Uddin et al.[8] also reported a 5-year-old female, with global DD and a genetic deletion that spanned BP4-BP5 encompassing *OTUD7A*, but not *CHRNA7*, as well as an ASD male individual carrying an inframe deletion in *OTUD7A* (NM_130901.2(*OTUD7A*):c.1474_1482del, p.(Asn492_Lys494del)). The inframe deletion was not inherited from the healthy parents and was also present in the affected brother.

These recent discoveries have drawn attention to *OTUD7A* as a major cause of the neurodevelopmental features associated with 15q13.3 microdeletion syndrome. The analysis of the mutational status of this gene in cohorts of unsolved patients presenting with DD/ID is thus justified. Furthermore, the physiopathological events underlying 15q13.3 microdeletion syndrome remain poorly characterized and, although the molecular mechanisms associated with OTUD7A dysfunction are starting to emerge,[8,17] further investigations are required in order to fully understand the processes underlying its implication in human disease.

Here, we report the first case of a homozygous missense variant (NM_130901.2:c.697C>T, p.(Leu233Phe)) in *OTUD7A* and related proteasome dysfunction found in a patient with severe global developmental delay, language impairment and epileptic encephalopathy. Based on our finding and recent literature,[8,17] this gene variant may

have a significant role in chromosome 15q13.3 microdeletion syndrome, associated with a wide range of neurodevelopmental disorders.

2 | METHODS

2.1 | Sample collection

Peripheral blood samples from the affected patient and his parents and a buccal swab from the patient's brother were obtained. Fibroblasts were derived from skin punch biopsies as previously described in Supporting Information and methods and obtained from the affected patient and from four unrelated age-matched healthy control individuals.

2.2 | DNA extraction, exome sequencing and validation of an identified variant

Genomic DNA extraction from blood was performed using the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) as recommended by the provider. Genomic DNA was extracted from the buccal swab using a standard phenol-chloroform procedure. Exome capture and sequencing and the resulting libraries were sequenced using the methods previously described in Nambot et al.[20] Complete exome sequencing (ES) section can be found in the Supporting Information and methods. The identified variant was validated by Sanger sequencing with forward 5′-TGTTTCTGCCTTCCCCTCAT-3′ and reverse 5′-TGGGAAGCCTCAGGAAGAAG-3′ polymerase chain reaction (PCR) primers.

2.3 | Commercial parental (wild-type) and *OTUD7A* knockout HAP1 haploid cell lines

Parental and CRISPR/Cas9-edited *OTUD7A* knockout (*OTUD7A*$^{-}$) HAP1 haploid cell lines were purchased from Horizon Discovery. The presence of the knockout variant (NM_130901.1:c.494_509del, p.Ser165Thrfs*2) was verified by Sanger sequencing with forward 5′-TAGAATCCTAACTGAAAACTGCCCA-3′, reverse 5′-TAGAATCCTAACTGAAAACTGCCCA-3′ PCR primers and sequencing primer 5′-CTTGGGGCTATCTTTCCTTCTCC-3′.

2.4 | Cell cultures

Fibroblasts were cultured in Dulbecco's Modified Eagle Medium (DMEM) high-glucose medium (HyClone Thermo Scientific) supplemented with 10% fetal bovine serum (FBS) and 1% Zell Schield (commercial mixture of antibiotics and anti-mycoplasma reagents [Minerva, Biovalley, France]). Cells were cultured at 37°C in a humidified 5% CO_2 atmosphere.

HAP1 cells were cultured in Iscove's modified Dulbecco's Medium supplemented with 10% FBS and 1% penicillin/streptomycin. Cells were cultured at 37°C in a humidified 5% CO_2 atmosphere.

2.5 | Western blot and fluorogenic peptide hydrolyzation assay

Western blot and fluorogenic peptide hydrolyzation assays were performed on control and patient fibroblasts as previously described in Poli et al[21] and Santiago-Sim et al,[19] respectively (see Supporting Information and methods).

2.6 | Ethics statement

Informed written consent was obtained from the parents for themselves and their sons, as well as from families of control donors.

The study was performed within the framework of the GAD ("Génétique des Anomalies du Développement") collection and approved by the appropriate institutional review board (DC2011-1332).

3 | RESULTS

3.1 | Case report

The proband was born at term after an uneventful pregnancy and delivery. His growth parameters were normal. He was the second male child of distant consanguineous Portuguese parents, both presenting with a learning disability. They had completed middle school but did not obtain a high school diploma. His 5-year-old brother was attending a regular school and presented nonspecific learning difficulties. However, his formal evaluation could not be performed because of the family context. Otherwise, the family history was unremarkable.

The initial evaluation of the patient occurred with the development of abnormal movements at 4 months described by the parents as loss of head control for few seconds in association with hypotonia. However, brain magnetic resonance imaging (MRI), performed at 5 months, was normal. Head control was acquired at 10 months. He was able to roll over by age 12 months. At 14 months, the electroencephalogram (EEG) highlighted a typical pattern of infantile spasms with hypsarrythmia (Figure S2A). The diagnosis of infantile spasms led to treatment with vigabatrin (VGB) and prednisolone, which resulted in complete control of the spasms 14 days after the initiation of treatment. Repeated EEGs showed a continuous sleep and awake abnormal background pattern with frequent, multifocal sharp waves. The corticosteroid therapy was progressively lowered after 1 month. A follow-up brain MRI at 1 year and 3 months was performed, corresponding to 1 month after the beginning of vigabatrin treatment. The imaging revealed unspecific anomalies: T2/FLAIR showed bilateral hyperintensity of the globi pallidi and of the ventroposterolateral

nuclei of the thalamus, enlarged subarachnoid space, in particular in the frontoparietal region, bilateral sulcal widening of the cerebral hemispheres, enlarged cerebellar interfolial spaces and a thin corpus callosum (Figure S3A). VGB-associated reversible MRI signal changes compatible with those presented by our patient have been described in the literature.[22] Proton MR Spectroscopy was normal. Lactic and pyruvic acid were normal in cerebrospinal fluid. Because of the control of the spasms, vigabatrin treatment was stopped at 18 months. The follow-up EEG performed 2 weeks after VGB withdrawal confirmed that the spasms were completely resolved and the patient was seizure-free (Figure S2B).

On the last exam at 28 months, the parents declared that he had been able to sit without support since the age of 23 months. He could not walk but was able to stand for a few seconds. He used three words: papa, maman, Kevin. He did not eat alone, did not point, and did not grasp with his whole hand. He was able to wave goodbye and because he had developed a pincer grasp, he was partially able to grab and hold objects and transfer across the midline. He also babbled, made eye contact and smiled in response to social stimuli. His weight and head circumference were 12.05 kg (−0.5 SD) and 48.5 cm (−2 SD), respectively. He presented hypotonia. Tendon reflexes were normal, and there were neither Babinski nor Hoffman signs, and no tremors or hypokinesia. He was seizure free until 28 months when he started to present atypical absence seizures characterized at the EEG by generalized spike and wave with motor manifestations (Figure S2C). He was thus initially treated with levetiracetam alone, allowing a significant decrease of seizures, which was subsequently associated with zonisamide allowing to achieve a seizure-free state.

3.2 | Outcome of exome sequencing of the patient

The family was referred to a clinical genetics unit, which requested tES and array-CGH of the patient in order to identify the genetics underlying the clinical manifestations.

A homozygous missense variant, chr15:g.31819467G>A (NM_130901.2:c.697C>T, p.(Leu233Phe)), was found in the *OTUD7A* gene (MIM: 612024; Figure S3B). Independent Sanger sequencing confirmed the presence of the variant at the homozygous state in the proband and the heterozygous carrier status of the parents. The extended segregation analysis revealed the presence of the variant at the heterozygous state in the brother as well (Figure S3C).

This variant was absent in the gnomAD database, and bioinformatics scores were in favor of a predicted pathogenic effect (GERP = 5.6, Polyphen2 = 1.0, MISZ = 5.917 and LoF intolerant [pLI] = 0.975). Functional prediction scores SIFT and Provean (on VarSome)[23] predicted that this variant would have a damaging impact.

The variant is located in the OTU-like cysteine protease catalytic domain and affects a leucine which is part of an alpha helix conserved through evolution and OTU A20-like subfamily (Figure S4).[24]

3.3 | OTUD7A dysfunction results in decreased levels of PA28-regulator and PA28-20S proteasome complexes formation in human fibroblasts

Given the potential implication of the OTU deubiquitinase family in the regulation of proteasome populations and activities,[19] we explored the impact of OTUD7A defects on the intracellular pools of 20S and 26S proteasome complexes either in fibroblasts derived form a healthy subject (CTRL1) or our patient carrying the homozygous missense c.697C>T variant. As pathogenic 15q13.3 microdeletions are associated with incomplete penetrance and highly variable expressivity at the heterozygous state,[2,3,5] the use of material derived from heterozygous c.697C>T carriers were not necessary for our experiments, because such material would not be necessarily informative. Whole-cell extracts were separated by native-PAGE with proteasome bands being visualized by their ability to hydrolyze the Suc-LLVY-AMC fluorogenic peptide. As expected, our in-gel proteasome activity assay revealed two strongly fluorescent bands corresponding to the 20S and 26S/hybrid proteasome complexes in these cells (Figure 1A). Interestingly, the enzymatic activity of the 20S complex in the *OTUD7A* mutant cells was significantly decreased when compared to that of the unrelated control fibroblasts. Direct measurement of the chymotrypsin-like activity in whole-cell lysates confirmed that this peptide hydrolyzing activity significantly declined in the *OTUD7A* subject (Figure 1B). The reduced 20S proteasome activity detected in the *OTUD7A* subject was not due to reduced amount of 20S proteasomes, as evidenced by western blotting using anti-α6 antibody (Figure 1A). Similarly, no major differences could be detected in the expression of the Rpn5 subunit between the two cell lines, confirming that the two fibroblast cell lines were endowed with similar amounts of 26S proteasome complexes. Interestingly, the anti-PA28-α antibody strongly stained two major bands corresponding to the positions of the 20S proteasome and free PA28 complex (~180 kDa) in the control fibroblasts, indicating that a substantial part of the 20S complex is associated with at least one PA28 ring in these cells. Strikingly, both PA28-20S and free PA28 complexes were found to be significantly decreased in the *OTUD7A* subject when compared to the control cell line. These data strongly suggest that a homozygous c.697C>T *OTUD7A* variant causes a drop in PA28 complexes. To confirm the loss of PA28 in the *OTUD7A* mutant cells, whole-cell extracts were also analyzed by sodium dodecyl sulfate - polyacrylamide gel electrophoresis followed by western blotting using antibodies directed against various proteasome subunits. As shown in Figure 1C and Figure S5, the steady-state expression levels of PA28-α and 20S proteasome subunits (ie, α2, β1, β2, β5, β1i, β2i and β5i) were severely compromised in OTUD7A mutant fibroblasts, while most of the other 20S (ie, α6) and 19S proteasome subunits (ie, Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6, Rpn5 and Rpn10) remain unaffected, when compared to control cell lines. The reduced proteasome activity detected in the OTUD7A subject was accompanied by a substantial accumulation of proteins modified with K48-linked ubiquitin chains which represent typical targets for proteasome-mediated degradation (Figure 1C). Altogether, these data point to a causal relationship between

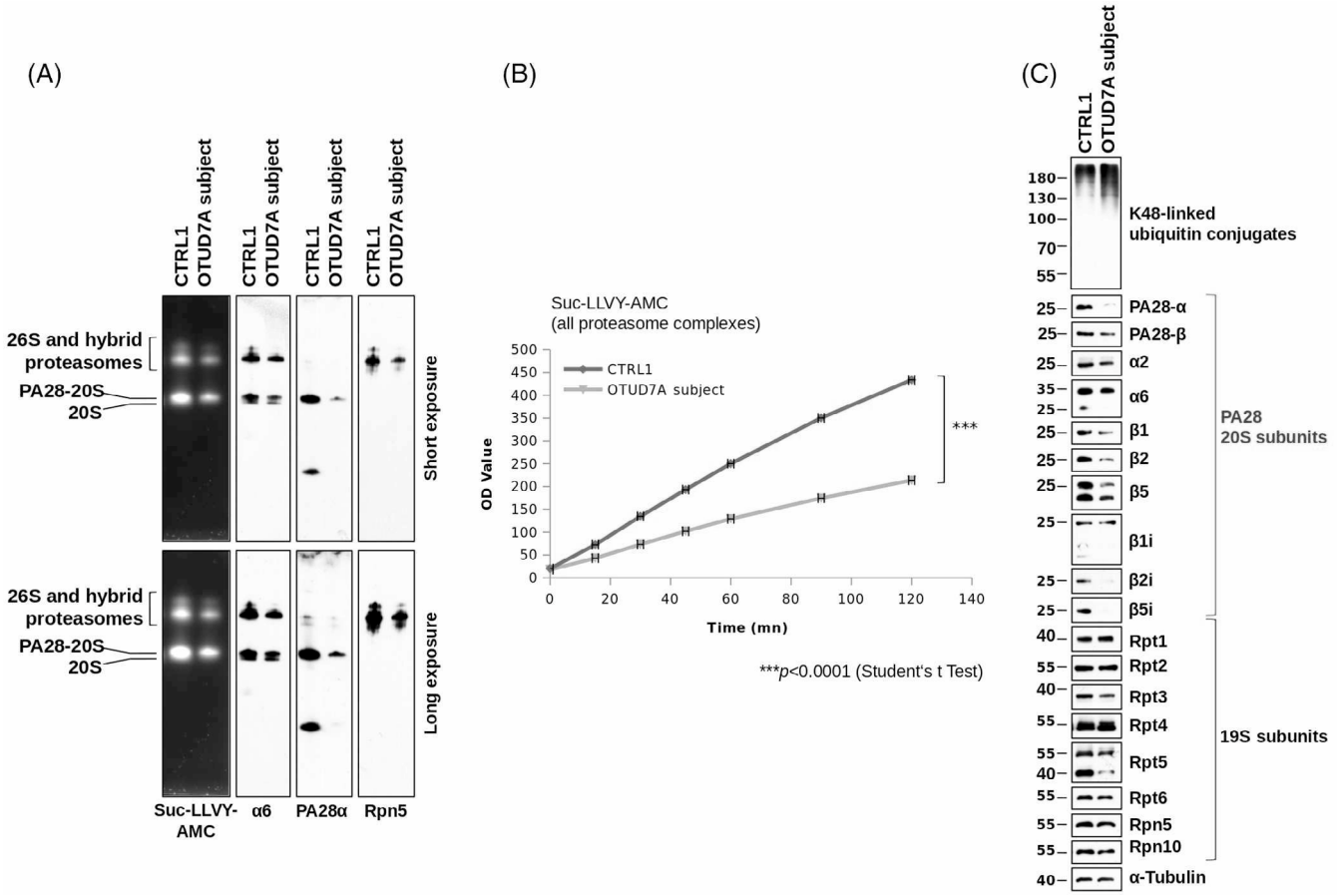


FIGURE 1 Fibroblasts derived from a subject carrying a homozygous missense c.697C>T mutation in *OTUD7A* fail to assemble active PA28-20 proteasome complexes. A, Whole-cell extracts from control (CTRL1) and OTUD7A-defective fibroblasts were separated by native PAGE and chymotrypsin-like activity was examined using 0.1 mM Suc-LLVY-AMC. In addition, the native-PAGE gels were subjected to western blotting using antibodies to α6, PA28-α and Rpn5, as indicated. Two exposure times are shown. B, Ten micrograms of whole-cell lysate from control (CTRL1) and *OTUD7A*-mutant fibroblasts were incubated in a final 100 μL volume containing 0.1 mM Suc-LLVY-AMC (in quadruplicates) on a 96-well plate at 37°C and subsequently subjected onto a Microplate Reader to measure fluorescence activity every 15 to 30 minutes for 2 hours. ***$P < .0001$ (Student *t* test). C, Protein extracts from control (CTRL1) and *OTUD7A*-mutant fibroblasts were prepared and subjected to SDS-PAGE followed by western blotting with antibodies specific for K48-linked ubiquitin chains, Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6, Rpn5, Rpn10, α6, PA28-α, β1, β2, β5, β1i, β2i, β5i and α-Tubulin (loading control) [Colour figure can be viewed at wileyonlinelibrary.com]

OTUD7A disruption and impaired assembly of PA28-20S complexes as a result of PA28 and 20S proteasome subunit deficiency.

3.4 | Functional analysis of OTUD7A dysfunction in HAP1 haploid cell lines

To confirm the unexpected interplay between OTUD7A and proteasome expression observed in human fibroblasts, we next assessed the proteasome content and activities in HAP1 haploid cells carrying a 16 bp deletion in the exon 3 of *OTUD7A* (NM_130901.1: c.494_509del, p.Ser165Thrfs*2). As illustrated in Figure 2, HAP1 *OTUD7A*$^{-}$ cells exhibit reduced amounts of PA28/20S complexes as well as decreased chymotrypsin-like activity, as determined by in-gel and in-plate assays (Figure 2A,B) and as compared to parental HAP1 wild-type cells. In a similar manner to the *OTUD7A* mutant fibroblast cell line isolated from our patient, HAP1 *OTUD7A*$^{-}$ cells had decreased steady-state expression levels of PA28-α and 20S proteasome subunits, when compared to their wild-type counterparts (Figure 2C). As expected, the inability of HAP1 *OTUD7A*$^{-}$ cells to express sufficient levels of PA28-α and 20S proteasome subunits resulted heavily of perturbed protein homeostasis, as evidenced by the aggregation of insoluble proteins marked with K48-linked ubiquitin chains (Figure 2).

4 | DISCUSSION

15q13.3 microdeletion syndrome, caused by heterozygous loss of the critical region encompassing *CHRNA7* and *OTUD7A*, is associated with incomplete penetrance and highly variable expressivity.[2-5] Rare homozygous or compound heterozygous patients have been also reported,

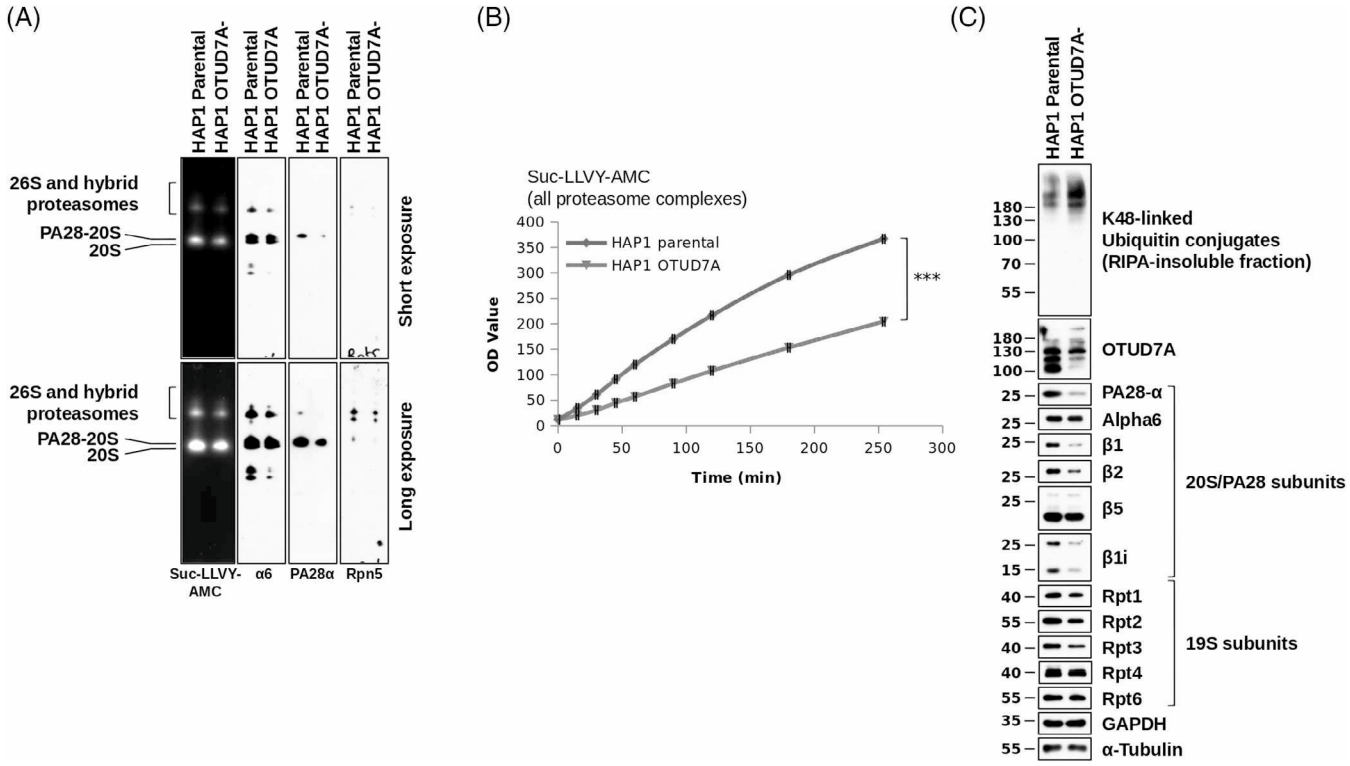


FIGURE 2 HAP1 cells deficient for *OTUD7A* show signs of perturbed protein homeostasis due to reduced proteasome activity as a result of impaired expression of PA28 and 20S proteasome subunits. A, HAP1 parental (wild-type) and devoid of OTUD7A (HAP *OTUD7A*$^{-}$) were subjected to protein extraction using a TSDG native lysis buffer (10 mM Tris pH 7.0, 10 mM NaCl, 25 mM KCl, 1.1 mM MgCl2, 0.1 mM EDTA, 1 mM DTT, 1 mM NaN3, 20% glycerol) and subsequently separated on 3% to 12% native PAGE for 16 hours at 4°C. Proteasome complexes were visualized by incubating the gels with 0.1 mM of the Suc-LLVY-AMC fluorogenic peptide using a Fusion FX imager (Vilber). The separated gels were further analyzed by western blotting using antibodies specific for α6, PA28-α and Rpn5, as indicated. B, Ten micrograms of the protein samples prepared in (A) were incubated in quadruplicates with 0.1 mM Suc-LLVY-AMC fluorogenic peptide in a final volume of 100 μL on 96-well plates and fluorescence was measured every 15 to 30 minutes over a 3-hours period of time using the exciting/emission filter set at 360/460 nm, respectively. C, Whole-cell lysates of HAP1 parental wild-type and *OTUD7A*$^{-}$ cells were prepared in RIPA buffer and separated by sodium dodecyl sulfate - polyacrylamide gel electrophoresis (SDS-PAGE) prior to western blotting analysis using antibodies directed to OTUD7A, PA28-α as well as various proteasome subunits including α6, β1, β2, β5, β1i, Rpt1, Rpt2, Rpt3, Rpt4, Rpt5 and Rpt6, as indicated. Membranes were further probed with antibodies specific for GAPDH and β-Tubulin to ensure equal protein loading. For detection of proteins modified with K48-linked ubiquitin chains, the RIPA-insoluble pellets of HAP1 parental and *OTUD7A*$^{-}$ samples were resuspended in urea lysis buffer (8M urea, 2M, thiourea, 4% CHAPS) prior to SDS-PAGE and western blotting [Colour figure can be viewed at wileyonlinelibrary.com]

but they present more severe phenotypes including epileptic encephalopathy, hypotonia and poor growth.[1,3,5-8] Although pathogenic microdeletions have been reported independently by several laboratories,[2-5] the contribution of these two genes in respect to the physio-pathological mechanisms underlying the neurodevelopmental phenotype associated with this syndrome has remained elusive.

However, recent investigations have provided insight into the contribution of the *OTUD7A* gene in the neurodevelopmental phenotype of patients with 15q13.3 microdeletion. This gene encodes for the OTU Domain-containing protein 7A belonging to a protein family involved in the regulation many cellular processes via deubiquitination, expressed in brain and localized to dendritic and synaptic compartments during the maturation of cortical neurons.[8] Ubiquitin is a conserved, globular protein consisting of 76 amino acids that can be attached to proteins either in a mono or polymerized form, impacting on their activity, localization, interactome, and turnover.[25] Ubiquitination is a reversible posttranslational modification, mediated by deubiquitinating enzymes, playing also a critical role in synapse formation and function.[26] Pathogenic variants in deubiquitinases such as USP9X, USP27X and OTUD6B have been already associated with recessive human diseases responsible for mental retardation (MIM 300919, 300 984 and 617 452, respectively).[19,27,28]

In 2018, a first report detailing two siblings with ASD and a de novo 9 bp in frame *OTUD7A* deletion (p.Asn492_Lys494del) and another patient with global developmental delay and a BP4-BP5 microdeletion including *OTUD7A* but not *CHRNA7* and its upstream region was published.[8] Null *Otud7a* mice[17] displayed growth delay, delayed motor milestone, seizure-like activity, impaired vocalization, significant reduction of dendritic spine density as well as decreased functioning excitatory synapses which can be rescued by wild-type OTUD7A expression.

Interestingly, Uddin et al.[8] also observed that reported cases carrying a minimal microdeletion in the region encompassing the single *CHRNA7* gene,[14,15] also included an intergenic region containing regulatory elements potentially accounting for *OTUD7A* expression. This suggests that the phenotypes observed in patients carrying *CHRNA7* deletions may not have resulted from CHRNA7 dysfunction, but instead from an aberrant *OTUD7A* expression and downstream activity of the encoded deubiquitinase.

The present study reports the first patient found to carry a homozygous pathogenic single nucleotide variant (NM_130901.2: c.697C>T, p.Leu233Phe) in *OTUD7A*. The patient presented with severe global developmental delay, language impairment, infantile spasms that progressed to atypical absence seizures. This clinical symptomatology is consistent with the phenotype observed in rare patients carrying homozygous or compound heterozygous 15q13.3 microdeletions. Abnormal brain findings were also highlighted in a follow-up MRI, while the patient was on Vigabatrin therapy. VGB-associated reversible MRI signal changes compatible with those presented by our patient have been described in the literature.[22] Unfortunately, the family did not return for a further follow-up MRI visits to determine whether any improvement had occurred, as could have been expected based on earlier reports.[22] However, MRI findings do not represent a constant trait associated with biallelic 15q13.3 variants. Interestingly, Simon et al.[6] have recently identified a patient with a homozygous 15q13.3 microdeletion encompassing *TRPM1*, *OTUD7A* and *CHRNA7* in a patient presenting with hypotonia, developmental delay, impaired vision, EEG findings and normal brain MRI results. The identification of additional patients carrying homozygous or compound heterozygous will be necessary to better characterize the brain MRI features of these patients.

No additional variants could account for the clinical presentation of the patient. The genetic tests performed, including array-CGH and tES also revealed the presence of a paternally inherited heterozygous deletion of 245 kb at 14q12 (arr[GRCh37] 14q12 (31904160_32148707)x1) encompassing three genes including *NUBPL*, and a rare maternally inherited variant (NM_022977.2: c.1657A>G, p.Thr553Ala) in *ACSL4*. However, these variants could not be considered as the primary cause of the clinical presentation of the patient, because either of an incompatible mode of inheritance or an inconsistent phenotype. Indeed, homozygous or compound heterozygous pathogenic variants in *NUBPL* are responsible for autosomal recessive mitochondrial complex I deficiency (OMIM: 613621).[29,30] Heterozygous partial deletions and loss-of-function variants of this gene have been identified in the healthy population, as reported in the Database of Genomic Variants (http://dgv.tcag.ca)[31] and in the Genome Aggregation Database (gnomAD—http://gnomad.broadinstitute.org),[32] respectively. The associated probability of being pLI for *NUBPL* is equal to 0 and it owns an observed/expected score of 0.53 with an associated 90% confidence interval of 0.35 to 0.93, suggesting that heterozygous pathogenic variants are frequent in the general population. As tES analysis did not identify any additional *NUBPL* variants, the inherited deletion was not considered responsible for the patient's phenotype. Similarly, pathogenic variants in *ACSL4* are known to cause nonspecific mental retardation (*MIM 300387*) without macroscopic brain abnormalities or epilepsy.[33,34] Based on the knowledge available in the literature, we did not consider *ACSL4* to be responsible for the epileptic encephalopathy or the anatomical brain abnormalities occurring in our patient, but rather a variant of unknown significance possibly contributing to ID (Supporting Information).

To gain insight into the molecular mechanisms underlying *OTUD7A* mutation, we performed functional investigations that allowed us to unveil an unexpected relationship between this gene and proteasome function. We discovered that OTUD7A regulates the steady-state expression level of the PA28 proteasome regulator and α- and β-subunits. In cells, protein degradation is achieved by two systems: the ubiquitin-proteasome system and the autophagy-lysosome system (reviewed in Rousseau and Bertolotti).[35] Proteins are targeted to the proteasome either through ubiquitination, usually in the form of lysine 48-linked polyubiquitin chains, or by the presence of an unstructured protein regions.[35] The eukaryotic 26S proteasome is a complex that consists of two different subcomplexes: the 20S core particle and the 19S regulatory particle (RP).[35] Additional regulatory complexes including PA28 composed by heteroheptamers can substitute the RP to assemble alternative forms of the proteasome α and β subunits.[35]

Our data show that fibroblasts carrying the homozygous c.697C>T missense variant in the *OTUD7A* gene exhibit lower protein levels of PA28 and α/β proteasome subunits than their wild-type counterparts (Figure 1 and Figure S5). Accordingly, this was accompanied by decreased formation of active PA28-20S proteasome complexes (Figure 1). Interestingly, such PA28-20S complexes have been shown to eliminate oxidant-damaged proteins in an ubiquitin-independent fashion.[36,37] Thus, it is also conceivable that OTUD7A-defective cells accumulate insoluble protein aggregates as a consequence of impaired breakdown of oxidized proteins. This hypothesis is in line with the recent reports showing that oxidative stress actively contributes to neurodevelopmental and/or neurodegenerative disorders.[38-40]

Strikingly, knocking out *OTUD7A* in the HAP1 haploid cell line reduced assembly of PA28-20S complexes and increased accumulation of ubiquitin-modified proteins which successfully replicates the proteasome phenotype initially observed in the patient-derived *OTUD7A* mutant cells (Figure S5). These data therefore unambiguously show a cause-and-effect relationship between OTUD7A deficiency and proteasome impairment. How OTUD7A influences the PA28 and α- and β-proteasome subunit steady-state expression levels remain unclear and further studies are required to determine whether the decreased amounts of PA28 and α/β subunits detected in OTUD7A defective cells are due to decreased transcription/translation efficiency or increased protein turnover.

Our work reinforces the notion that OTU deubiquitinases actively participate in the regulation of proteasome assembly and/or the abundance of proteasome isoforms.[19] Indeed, it was previously shown that cells from subjects suffering from cognitive impairment and carrying biallelic deletions in the *OTUD6B* gene fail to incorporate 19S subunits

into 26S proteasomes.[19] Here, we identify *OTUD7A* as a new member of the growing family of deubiquitinases whose functional disruption has an impact on the intracellular proteasome pools. Interestingly and unlike OTUD6B, OTUD7A seems to be required for PA28 expression, even though the underlying mechanism remains ill-defined. Given that OTUD7A has been shown to specifically remove ubiquitin chains branched at lysine (K)11,[18,24] we hypothesize that the maintenance of PA28 expression levels involves the cleavage of K11-linked ubiquitin chains from PA28 itself and/or other regulators. In this regard, further investigations will be needed to address the ubiquitination profile of PA28 as well as its potential impact on PA28 protein abundance.

Our observations provide additional evidence regarding the role of *OTUD7A* in neurocognitive conditions. We describe here a unique patient with a homozygous missense variant in the *OTUD7A* gene and severe epileptic encephalopathy. His phenotype was similar to previous reports of homozygous/compound heterozygous 15q13.3 microdeletion. Given this data and results of recent studies, *OTUD7A* gene could be considered a strong candidate for the neurocognitive phenotype of patients with the 15q13.3 deletion syndrome and for early-onset epileptic encephalopathies in a homozygous/compound heterozygous state.

ACKNOWLEDGEMENTS

This work was supported by grants from the Regional Council of Burgundy (to C.T.-R.), the FEDER 2017, PARI 2017 and CIFRE (ANRT) between Laboratoire Cerba and the Regional Council of Burgundy for the doctoral work at Laboratoire Cerba and GAD as well as by grants from the German Research Foundation (SFBTR 167 to E.K.), the Fritz-Thyssen Foundation (Az: 10.16.2.022MN to E.K.) and the Molecular Medicine Research Consortium of the University of Greifswald (FOVB-2018-11 to F.E.). We are grateful to Robert Beyer for excellent technical assistance and thank the family for their participation. We would like to thank Suzanne Rankin (CHU Dijon Bourgogne) for proofreading the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

For the functional section, the data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Philippine Garret https://orcid.org/0000-0003-2551-3606
Alain Verloes https://orcid.org/0000-0003-4819-0264
Antonio Vitobello https://orcid.org/0000-0003-3717-8374

REFERENCES

1. Endris V, Hackmann K, Neuhann TM, et al. Homozygous loss of CHRNA7 on chromosome 15q13.3 causes severe encephalopathy with seizures and hypotonia. *Am J Med Genet A*. 2010;152A(11):2908-2911. https://doi.org/10.1002/ajmg.a.33692.
2. Sharp AJ, Mefford HC, Li K, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet*. 2008;40(3):322-328. https://doi.org/10.1038/ng.93.
3. Masurel-Paulet A, Andrieux J, Callier P, et al. Delineation of 15q13.3 microdeletions. *Clin Genet*. 2010;78(2):149-161. https://doi.org/10.1111/j.1399-0004.2010.01374.x.
4. Masurel-Paulet A, Drumare I, Holder M, et al. Further delineation of eye manifestations in homozygous 15q13.3 microdeletions including TRPM1: a differential diagnosis of ceroid lipofuscinosis. *Am J Med Genet A*. 2014;164(6):1537-1544. https://doi.org/10.1002/ajmg.a.36471.
5. Lowther C, Costain G, Stavropoulos DJ, et al. Delineating the 15q13.3 microdeletion phenotype: a case series and comprehensive review of the literature. *Genet Med*. 2015;17(2):149-157. https://doi.org/10.1038/gim.2014.83.
6. Simon J, Stoll K, Fick R, Mott J, Lawson-Yuen A. Homozygous 15q13.3 microdeletion in a child with hypotonia and impaired vision: a new report and review of the literature. *Clin Case Rep*. 2019;7(12):2311-2315. https://doi.org/10.1002/ccr3.2403.
7. Spielmann M, Reichelt G, Hertzberg C, et al. Homozygous deletion of chromosome 15q13.3 including CHRNA7 causes severe mental retardation, seizures, muscular hypotonia, and the loss of KLF13 and TRPM1 potentially cause macrocytosis and congenital retinal dysfunction in siblings. *Eur J Med Genet*. 2011;54(4):e441-e445. https://doi.org/10.1016/j.ejmg.2011.04.004.
8. Uddin M, Unda BK, Kwan V, et al. OTUD7A regulates neurodevelopmental phenotypes in the 15q13.3 microdeletion syndrome. *Am J Hum Genet*. 2018;102(2):278-295. https://doi.org/10.1016/j.ajhg.2018.01.006.
9. Nilsson SRO, Celada P, Fejgin K, et al. A mouse model of the 15q13.3 microdeletion syndrome shows prefrontal neurophysiological dysfunctions and attentional impairment. *Psychopharmacology (Berl)*. 2016;233:2151-2163. https://doi.org/10.1007/s00213-016-4265-2.
10. Forsingdal A, Fejgin K, Nielsen V, Werge T, Nielsen J. 15q13.3 homozygous knockout mouse model display epilepsy-, autism- and schizophrenia-related phenotypes. *Transl Psychiatry*. 2016;6(7):e860. https://doi.org/10.1038/tp.2016.125.
11. Soler-Alfonso C, Carvalho CM, Ge J, et al. CHRNA7 triplication associated with cognitive impairment and neuropsychiatric phenotypes in a three-generation pedigree. *Eur J Hum Genet*. 2014;22(9):1071-1076. https://doi.org/10.1038/ejhg.2013.302.
12. Hasselmo ME. The role of acetylcholine in learning and memory. *Curr Opin Neurobiol*. 2006;16:710-715. https://doi.org/10.1016/j.conb.2006.09.002.
13. Levin ED. α7-Nicotinic receptors and cognition. *Curr Drug Targets*. 2012;13(5):602-606. https://doi.org/10.2174/138945012800398937.
14. Hoppman-Chaney N, Wain K, Seger P, Superneau D, Hodge J. Identification of single gene deletions at 15q13.3: further evidence that CHRNA7 causes the 15q13.3 microdeletion syndrome phenotype: identification of CHRNA7 deletions at 15q13.3. *Clin Genet*. 2013;83(4):345-351. https://doi.org/10.1111/j.1399-0004.2012.01925.x.
15. Prasun P, Hankerd M, Kristofice M, Scussel L, Sivaswamy L, Ebrahim S. Compound heterozygous microdeletion of chromosome 15q13.3 region in a child with hypotonia, impaired vision, and global developmental delay. *Am J Med Genet A*. 2014;164(7):1815-1820. https://doi.org/10.1002/ajmg.a.36535.
16. Yin J, Chen W, Yang H, Xue M, Schaaf CP. Chrna7 deficient mice manifest no consistent neuropsychiatric and behavioral phenotypes. *Sci Rep*. 2017;7:39941. https://doi.org/10.1038/srep39941.
17. Yin J, Chen W, Chao ES, et al. Otud7a knockout mice recapitulate many neurological features of 15q13.3 microdeletion syndrome. *Am J Hum Genet*. 2018;102(2):296-308. https://doi.org/10.1016/j.ajhg.2018.01.005.

18. Mevissen TET, Kulathu Y, Mulder MPC, et al. Molecular basis of Lys11-polyubiquitin specificity in the deubiquitinase Cezanne. *Nature*. 2016;538(7625):402-405. https://doi.org/10.1038/nature19836.
19. Santiago-Sim T, Burrage LC, Ebstein F, et al. Biallelic variants in OTUD6B cause an intellectual disability syndrome associated with seizures and Dysmorphic features. *Am J Hum Genet*. 2017;100(4):676-688. https://doi.org/10.1016/j.ajhg.2017.03.001.
20. Nambot S, Thevenon J, Kuentz P, et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet Med*. 2018;20(6):645-654. https://doi.org/10.1038/gim.2017.162.
21. Poli MC, Ebstein F, Nicholas SK, et al. Heterozygous truncating variants in POMP escape nonsense-mediated decay and cause a unique immune dysregulatory syndrome. *Am J Hum Genet*. 2018;102(6):1126-1142. https://doi.org/10.1016/j.ajhg.2018.04.010.
22. Dracopoulos A, Widjaja E, Raybaud C, Westall CA, Snead OC. Vigabatrin-associated reversible MRI signal changes in patients with infantile spasms. *Epilepsia*. 2010;51(7):1297-1304. https://doi.org/10.1111/j.1528-1167.2010.02564.x.
23. Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978-1980. https://doi.org/10.1093/bioinformatics/bty897.
24. Mevissen TET, Hospenthal MK, Geurink PP, et al. OTU deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. *Cell*. 2013;154(1):169-184. https://doi.org/10.1016/j.cell.2013.05.046.
25. Pinto-Fernández A, Davis S, Schofield AB, et al. Comprehensive landscape of active deubiquitinating enzymes profiled by advanced chemoproteomics. *Front Chem*. 2019;7:592. https://doi.org/10.3389/fchem.2019.00592.
26. Kowalski JR, Juo P. The role of deubiquitinating enzymes in synaptic function and nervous system diseases. *Neural Plast*. 2012;2012:1-13. https://doi.org/10.1155/2012/892749.
27. Homan CC, Kumar R, Nguyen LS, et al. Mutations in USP9X are associated with X-linked intellectual disability and disrupt neuronal cell migration and growth. *Am J Hum Genet*. 2014;94(3):470-478. https://doi.org/10.1016/j.ajhg.2014.02.004.
28. Hu H, Haas SA, Chelly J, et al. X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Mol Psychiatry*. 2016;21(1):133-148. https://doi.org/10.1038/mp.2014.193.
29. Calvo SE, Tucker EJ, Compton AG, et al. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat Genet*. 2010;42(10):851-858. https://doi.org/10.1038/ng.659.
30. Kevelam SH, Rodenburg RJ, Wolf NI, et al. NUBPL mutations in patients with complex I deficiency and a distinct MRI pattern. *Neurology*. 2013;80(17):1577-1583. https://doi.org/10.1212/WNL.0b013e31828f1914.
31. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42:D986-D992. https://doi.org/10.1093/nar/gkt958.
32. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 2019;531210. https://doi.org/10.1101/531210.
33. Meloni I, Muscettola M, Raynaud M, et al. FACL4, encoding fatty acid-CoA ligase 4, is mutated in nonspecific X-linked mental retardation. *Nat Genet*. 2002;30(4):436-440. https://doi.org/10.1038/ng857.
34. Longo I, Frints SGM, Fryns J-P, et al. A third MRX family (MRX68) is the result of mutation in the long chain fatty acid-CoA ligase 4 (FACL4) gene: proposal of a rapid enzymatic assay for screening mentally retarded patients. *J Med Genet*. 2003;40(1):11-17. https://doi.org/10.1136/jmg.40.1.11.
35. Rousseau A, Bertolotti A. Regulation of proteasome assembly and activity in health and disease. *Nat Rev Mol Cell Biol*. 2018;19(11):697-712. https://doi.org/10.1038/s41580-018-0040-z.
36. Pickering AM, Koop AL, Teoh CY, Ermak G, Grune T, Davies KJA. The Immunoproteasome, the 20S proteasome, and the PA28αβ proteasome regulator are oxidative-stress-adaptive proteolytic complexes. *Biochem J*. 2010;432(3):585-594. https://doi.org/10.1042/BJ20100878.
37. Brehm A, Krüger E. Dysfunction in protein clearance by the proteasome: impact on autoinflammatory diseases. *Semin Immunopathol*. 2015;37(4):323-333. https://doi.org/10.1007/s00281-015-0486-4.
38. Reeg S, Grune T. Protein oxidation in aging: does it play a role in aging progression? *Antioxid Redox Signal*. 2015;23(3):239-255. https://doi.org/10.1089/ars.2014.6062.
39. Adav SS, Sze SK. Insight of brain degenerative protein modifications in the pathology of neurodegeneration and dementia by proteomic profiling. *Mol Brain*. 2016;9:92. https://doi.org/10.1186/s13041-016-0272-9.
40. Wells PG, Bhatia S, Drake DM, Miller-Pinsler L. Fetal oxidative stress mechanisms of neurodevelopmental deficits and exacerbation by ethanol and methamphetamine. *Birth Defects Res C Embryo Today*. 2016;108(2):108-130. https://doi.org/10.1002/bdrc.21134.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Garret P, Ebstein F, Delplancq G, et al. Report of the first patient with a homozygous *OTUD7A* variant responsible for epileptic encephalopathy and related proteasome dysfunction. *Clin Genet*. 2020;97:567–575. https://doi.org/10.1111/cge.13709

DEUXIÈME PARTIE : ANALYSE DE L'ADN MITOCHONDRIAL À PARTIR DE DONNÉES DE SÉQUENÇAGE D'EXOME

I- INTRODUCTION

L’analyse de l’ADNmt à partir de données de séquençage d’exome présente des défis à plusieurs niveaux. Le premier concerne le début de l’analyse bioinformatique que constitue l’étape d’alignement. En effet, la version actuelle de la référence mitochondriale rCRS (« revised Cambridge Reference Sequence » ou NC_012920.1 (Andrews et al., 1999)) n’est intégrée qu’à la version du génome de référence GRCh38, postérieure à GRCh37 utilisée en routine par de nombreuses équipes. Jusqu’à fin 2019, les principales bases de données de variations génomiques, dont la base gnomAD, utilisaient toujours la version GRCh37 du génome de référence. Les données de séquençage d’exome devaient donc être réalignées sur GRCh38 avant extraction de l’ADN mitochondrial en attendant une utilisation en routine de cette version du génome de référence. Une autre étape de l’analyse bioinformatique, que constitue l’appel de variant ou « variant calling », peut présenter des défis. En effet, les régions pseudomitochondriales ont des séquences similaires à celles de la mitochondrie. Si leur existence au sein des régions capturées permet d'accéder indirectement à la séquence de l'ADNmt, elles posent problème lors de l’analyse bioinformatique de l’ADNmt. Elles s’alignent à tort avec la référence mitochondriale. La mise en évidence de variations et la détermination du taux d’hétéroplasmie peuvent ainsi être biaisées. De plus, l’existence des haplogroupes mitochondriaux nécessite une étape supplémentaire d’analyse du génome mitochondrial car la détermination de l’haplogroupe des patients est nécessaire pour filtrer ces variations qui dépendent de l’origine des patients. Ainsi, le tri et l’annotation des variations mitochondriales diffèrent de ceux des variations nucléaires. Enfin, l’interprétation des données obtenues à l’issue de l’analyse bioinformatique est rendue plus complexe de par l’existence des taux d’hétéroplasmie qui varient entre les tissus, entre les individus et qui ont un impact sur le phénotype des patients.

II- MATÉRIEL ET MÉTHODES

II.1- Contrôles positifs et cohorte de patients

La méthode d’étude du génome mitochondrial à partir des données de séquençage à haut débit d’exome a été validée par l’analyse de 4 contrôles positifs porteurs d’une variation mitochondriale pathogène et bien caractérisée dans la littérature. L’exome de ces 4 individus a été séquencé après enrichissement avec le kit de capture Agilent CRE V2 ou Agilent v5.

Cette méthode a ensuite été appliquée sur une cohorte de 928 patients séquencés en exome dont 249 (26,8%) avaient un résultat d’exome nucléaire positif. Les patients atteints d’anomalies du développement avec ou sans atteinte neurologique représentaient ~80 % de la cohorte et ceux présentant une maladie neurologique rare ~20 %. La majorité des patients était d’origine caucasienne avec un âge variant de 0 (fœtus) à 87 ans. La proportion homme-femme était de 50 %.

Un séquençage à haut débit d'exome en solo ou trio à partir des échantillons ADN a été réalisé chez chaque patient. La capture a été effectuée avec un kit de chez Agilent : v3, v4, v5, v6, v7, Clinical Research Exome, Clinical Research Exome v2 ou plus récemment avec le kit TWIG. Ce dernier correspond au kit Human Core Exome de chez Twist Bioscience complété avec des sondes supplémentaires. Le séquençage a été réalisé soit sur un HiSeq 2000, HiSeq 4000, NextSeq 550 ou NovaSeq 6000 (Integragen) avec une lecture en paired-end et des lectures de 100 pb.

II.2- Détermination de l’haplogroupe

L’haplogroupe de chaque patient a été déterminé par HaploGrep 2.0 (Weissensteiner et al., 2016) à partir du fichier VCF contenant les variations mitochondriales brutes.

II.3- Base de données Mitomap

La base de données Mitomap (Lott et al., 2013) regroupe les données du séquençage de 50175 ADNmt humains complets et de 73 294 régions contrôles (mise à jour du 1er janvier 2020). Les variations identifiées y sont annotées avec la fréquence observée au sein de la base de données GenBank. Mitomap répertorie également les données des variations mitochondriales décrites dans la littérature. La pathogénicité des variations est indiquée par le statut « confirmé », « rapporté » ou « incertain ». Le premier statut est attribué aux variations dont la pathogénicité est admise la communauté scientifique et a été démontrée par au minimum deux études. Le statut de « rapporté » est attribué à une variation lorsqu’il existe une publication décrivant sa pathogénicité mais que d’autres études sont nécessaires pour renforcer cette conclusion. Enfin, le statut « incertain » concerne les variations pour lesquelles les résultats des études sont contradictoires. Pour les trois catégories le statut hétéroplasmique ou homoplasmique est également décrit.

II.4- Analyse des données d’exome indirectes

Le pipeline d’analyse du génome mitochondrial a été implémenté en suivant les recommandations du Broad Institute et en se basant sur le pipeline d’exome nucléaire utilisé en routine par l’équipe GAD (Thevenon et al., 2014). A l’inverse de l’exome nucléaire, les données brutes ont été alignées sur le génome de référence GRCh38 (**Fig. 36a**) car cette version du génome humain contient la référence actuelle de l’ADN mitochondrial GenBank NC_012920.1 (Andrews et al., 1999) contrairement à la référence GRCh37. De plus, le chromosome M est extrait par Picard Tools (Broad Institute) (v.2.4.1) à partir du fichier BAM brut avant les étapes de déduplication (Picard Tools) et de recalibration (GATK). L’alignement simultané des données d’exome sur les références nucléaire et mitochondriale est considéré comme la meilleure approche (Ye et al., 2014). Les variations mitochondriales ont été détectées à partir du fichier BAM recalibré, par l’outil HaplotypeCaller de GATK. L’ensemble des

résultats obtenus ont été réunis dans un même fichier VCF. Le premier filtre appliqué sur les variations brutes est celui de la fréquence. N'ont été conservées que les variations dont la fréquence est inférieure à 1 % (1000 Genomes Project Consortium et al., 2012) au sein de la base de données Mitomap. Les variations synonymes ont ensuite été filtrées (**Fig. 36b**). Les variations présentes chez plus de 4 individus n'ont également pas été retenues. En effet, la fréquence des variations pathogènes au sein de cette cohorte est inférieure à 3/928 patients. Ce seuil a donc été choisi pour filtrer les variations. Les variations impliquées dans un ou plusieurs haplogroupes (Van Oven, 2015) ont ensuite été annotées. Lorsque tous les individus porteurs d'une variation donnée possédaient un même haplogroupe défini par cette dernière, la variation a été considérée comme un polymorphisme et n'a pas été retenue. Le statut dans la base Mitomap (Lott et al., 2013) des variations restantes a ensuite été vérifié. Les variations considérées comme « confirmées », « rapportées » ou de « incertaines » ont été conservées pour la suite de l'analyse. Pour les autres variations absentes ou présentes sans statut, des bases de données publiques (ex. Clinvar (Landrum et al., 2018), OMIM (OMIM - Online Mendelian Inheritance in Man., 1996), etc.) ont été consultées ainsi que la littérature scientifique. Les phénotypes des individus ont été comparés à ceux décrits dans Mitomap ou la littérature. En cas de statut « confirmé » et d'une probable concordance phénotypique, la variation a été considérée comme candidate et validée par une deuxième méthode moléculaire. Il s'agit d'un score fourni par la base de données Mitomap qui prédit la probabilité pour une variation d'être pathogène. Un score inférieur à 25 % indique le caractère très probablement bénin de la variation, un score entre 25 et 50 % le caractère probablement bénin, un score entre 50 et 75 % un caractère probablement pathogène et un score supérieur à 75 % un caractère très probablement pathogène. En cas de non-concordance phénotypique ou d'absence dans les bases de données, les régions voisines ont été inspectées pour mettre en évidence d'éventuelles variations décrites chez des patients aux phénotypes similaires.

Afin de s'assurer que les variations mitochondriales candidates ne sont pas dues à l'existence de régions pseudomitochondriales au sein du génome nucléaire ces dernières ont été vérifiées. Les coordonnées des régions pseudomitochondriales équivalentes aux régions des variations mitochondriales candidates ont été extraites (Calabrese et al., 2012) et vérifiées

manuellement.

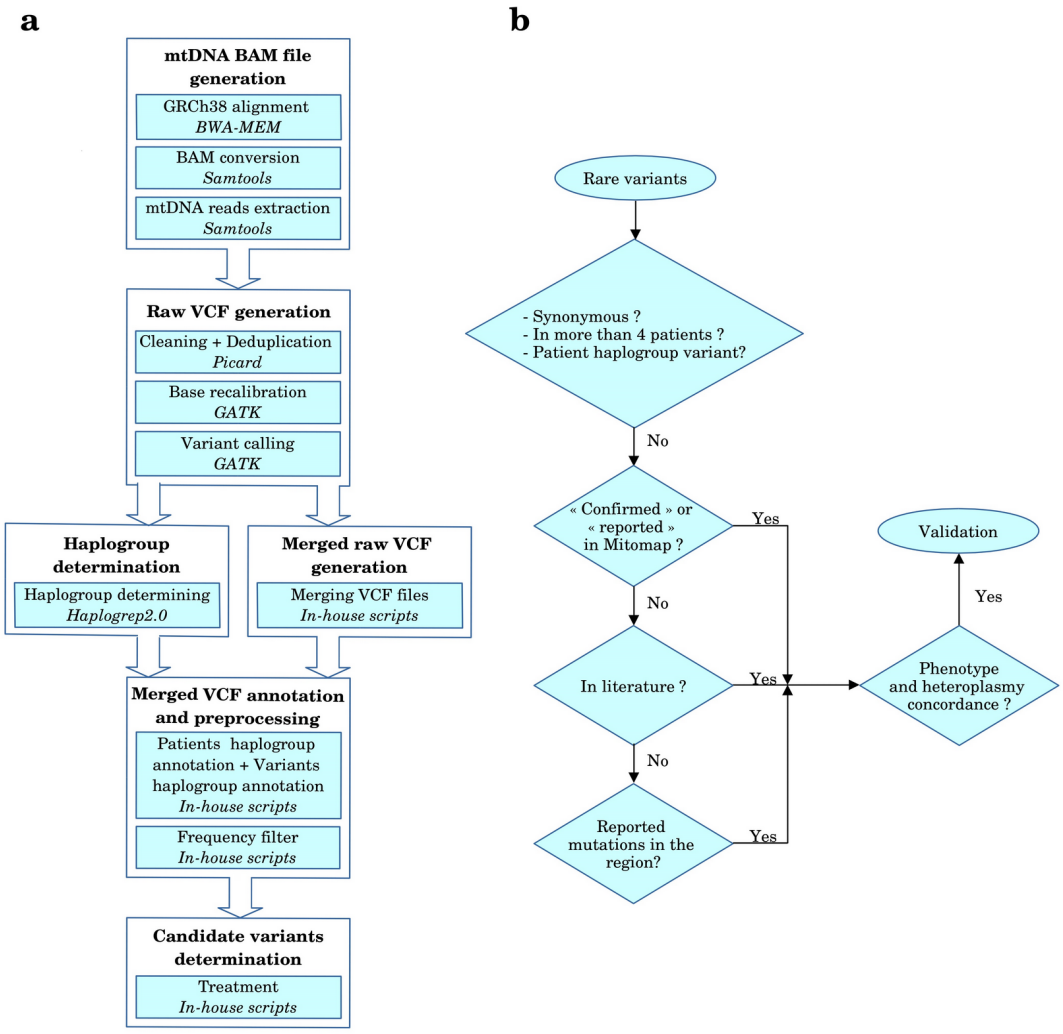


Figure 36 : Analyse de l'ADN mitochondrial à partir des données de séquençage à haut débit *a) Représentation du pipeline d'analyse de l'ADNmt à partir des données fastq jusqu'au VCF contenant les variations filtrées. L'obtention des fichiers BAM et des fichiers VCF bruts est basée sur le pipeline d'analyse de l'exome nucléaire utilisée en routine au laboratoire (Thevenon et al., 2016). La détermination de l'haplogroupe des patients a été réalisée par le programme HaploGrep 2.0 intégré au pipeline. Pour cette étude l'intégralité des variations détectées pour les 928 patients ont été regroupées dans un fichier VCF commun. L'haplogroupe des patients et les variations impliquées dans un ou plusieurs haplogroupes ont été annotés. La fréquence de chaque variation a été extraite et calculée à partir de la base de données GenBank présentes sur Mitomap. Les variations ayant une fréquence supérieure à 1 % au sein de la base de données Mitomap ont été filtrées. Les outils utilisés à chaque étape du pipeline sont indiqués en italique. b) Stratégies d'analyse des variations mitochondriales rares. Les variations synonymes, présentes chez plus de 4 individus de la cohorte et considérées comme impliquées dans l'haplogroupe ont été éliminées. Les variations « confirmées » ou « rapportées » dans Mitomap et/ou dans la littérature ont été étudiées.*

II.5- Validations par Sanger et par PCR-RFLP de variations mitochondriales

La validation des variations mitochondriales m.1494T>C, m.1555G>A et m.14484T>C a été réalisées par séquençage Sanger par l'équipe du Pr Procaccio. Les régions d'intérêt ont été amplifiées par PCR avec amorces spécifiques (**Tableau 4**) : une étape de 10 minutes à 94°C, 35 cycles d'une minute à 94°C suivie d'une minute à 60°C et d'une minute à 72°C, et une étape finale de 5 minutes à 72°C. Les amplicons obtenus à l'issue de la PCR ont été séquencés selon la méthode décrite dans Bonneau et al., (2014). La comparaison des séquences obtenues au génome mitochondrial de référence a été réalisée avec le programme Seqscape v2.7 (Thermofischer Scientific).

Variation	PCR Primers	
	Forward (5' → 3')	Reverse (5' → 3')
m.1494T>C	GAACACACAATAGCTAAGACCC	TTGGACAACCAGCTATCACCA
m.1555A>G	CCTCAAGTATACTTCAAAGGAC	GGCGATAGAAATTGAAACCTG
m.14484T>C	AACCCAAAAAGGCATAATTAAAC	GATATGAAAAACCATCGTTGTAT

Tableau 4 : Amorces spécifiques du séquençage Sanger des variations mitochondriales m.1494T>C, m.1555G>A et m.14484T>C

La PCR-RFLP a été utilisée, par l'équipe du Pr Procaccio, pour valider les variations m.3243A>G, m.8993T>G et m.11778G>A et calculer leur taux d'hétéroplasmie. La région d'intérêt a été amplifiée par PCR avec amorces spécifiques (**Tableau 5**) : une étape de 10 minutes à 94°C, 30 cycles d'une minute à 94°C suivie d'une minute à 58°C et d'une minute à 72°C, et une étape finale de 5 minutes à 72°C. Les amplicons obtenus ont ensuite été digérés, selon les instructions du fournisseur, par une enzyme de restriction spécifique de la variation à étudier. Les fragments obtenus après digestion ont été étudiés par électrophorèse capillaire (Applied 3130XL). Les résultats ont été analysés avec le programme Peak Scanner Analysis Software (Thermo Fisher Scientific). Le taux d'hétéroplasmie a été déterminé par le pourcentage de fragments porteurs de la variation.

Variation	PCR Primers		Restriction		
	Forward (5' → 3')	Reverse (5' → 3')	Enzyme	T° (°C)	Time (min)
m.3243A>G	*CCCTGTACGAAAGGACAAGAGAAATAACGCC	CGTTCGGTAAGCATTAGGAATGCCATTGC	HaeIII	37	30
m.8993T>G	AAATGCCCGAGCCCACTTCTTA	*GGTGGCGCTTCCAATTAGGT	BstNI	60	60
m.11778G>A	*GCCCACGGGCTTACATC	AAACCCGGTAATGATGTCGG	Tsp45I	65	120

*6-FAM Dye

Tableau 5 : Conditions des analyses par PCR-RFLP

II.6- Calcul du taux d’hétéroplasmie par NGS

Le taux d'hétéroplasmie des variations m.9035T>C, m.14502T>C et m.13051G>A a été réalisée par séquençage haut débit du génome mitochondrial. L’ADN mitochondrial a été amplifié en deux fragments chevauchants de 8009 et 8994 pb (Boucret et al., 2017) qui ont subi une fragmentation enzymatique. Les fragments ont été séquencés par Ion S5-XL (Thermo Fisher Scientific) selon les indications du fabricant. Le taux d’hétéroplasmie correspond au ratio du nombre de lectures porteuses de la variation divisé par le nombre total de lectures à cette position.

III- RÉSULTATS : MISE EN ÉVIDENCE DE VARIATIONS MITOCHONDRIALES PATHOGÈNES CAUSALES ET SECONDAIRES

L’étude des 4 contrôles positifs a montré que cette méthode d’analyse de l’ADN mitochondrial permettait bien de mettre en évidence les 4 variations causales pathogènes avec des taux d’hétéroplasmie similaires (**Tableau 6**). De plus, le coefficient de corrélation (Pearson)

entre les taux d'hétéroplasmie est de 0,943. Ce coefficient permet de détecter une relation linéaire entre ces deux séries de mesure (ici l'hétéroplasmie calculée lors de cette étude et l'hétéroplasmie déterminée par les méthodes ciblées). Un coefficient de valeur absolue proche de 1 montre qu'il existe une relation linéaire entre les deux séries. Le signe positif indique que les séries varient dans le même sens : si l'une augmente, l'autre augmente également.

Sexe	Age (années)	Variation mitochondriale	Hétéroplasmie attendue (méthode de détermination)	Pathologies associées aux variations	Hétéroplasmie détectée par le pipeline
M	14	m.3243A>G	72 % (PCR-RFLP)	MELAS/MIDD	73,7 %
M	32	m.3243A>G	21 % (PCR-RFLP)	MELAS/MIDD	50,0 %
M	25	m.13051G>A	86 % (NGS)	LHON	88 %
F	42	m.8993T>G	92-99 % (PCR-RFLP)	NARP/Leigh Disease / MILS	79,3 %

LHON : Leber's Hereditary Optic Neuropathy; MELAS: Mitochondrial Encephalomyopathy, Lactic Acidosis and Stroke-like episodes; MIDD: Maternally Inherited Diabetes and Deafness; MILS: Maternally Inherited Leigh Syndrome; NARP: Neurogenic muscle weakness, Ataxia, and Retinitis Pigmentosa. La séquence de référence est NC_012920.1.

Tableau 6 : Contrôles positifs porteurs de variations « confirmées » dans Mitomap et détectées par méthode ciblée en amont de cette étude

Avec les 928 patients de la cohorte, un total de 2429 variations brutes mitochondriales a été mis en évidence (**Fig. 37a**). En ne conservant que celles ayant une fréquence inférieure à 1 % dans la base de données GenBank sur Mitomap, 410 ont été éliminées. Sur les 2019/2429 restantes, presque la moitié n'étaient pas des variations synonymes et donc conservées pour la suite de l'analyse soit 1200/2019. Parmi elles, 984 étaient présentes chez moins de 4 individus. En retirant les variations définissant l'haplogroupe des patients concernés, 643/984 variations restaient à analyser. L'étude de la littérature et de la base de données Mitomap a permis de mettre en évidence 31/643 variations chez 39/928 patients (4,2 %). Ces candidats avaient un statut « confirmé », « rapporté » ou « incertain » dans Mitomap (**Fig. 37b**).

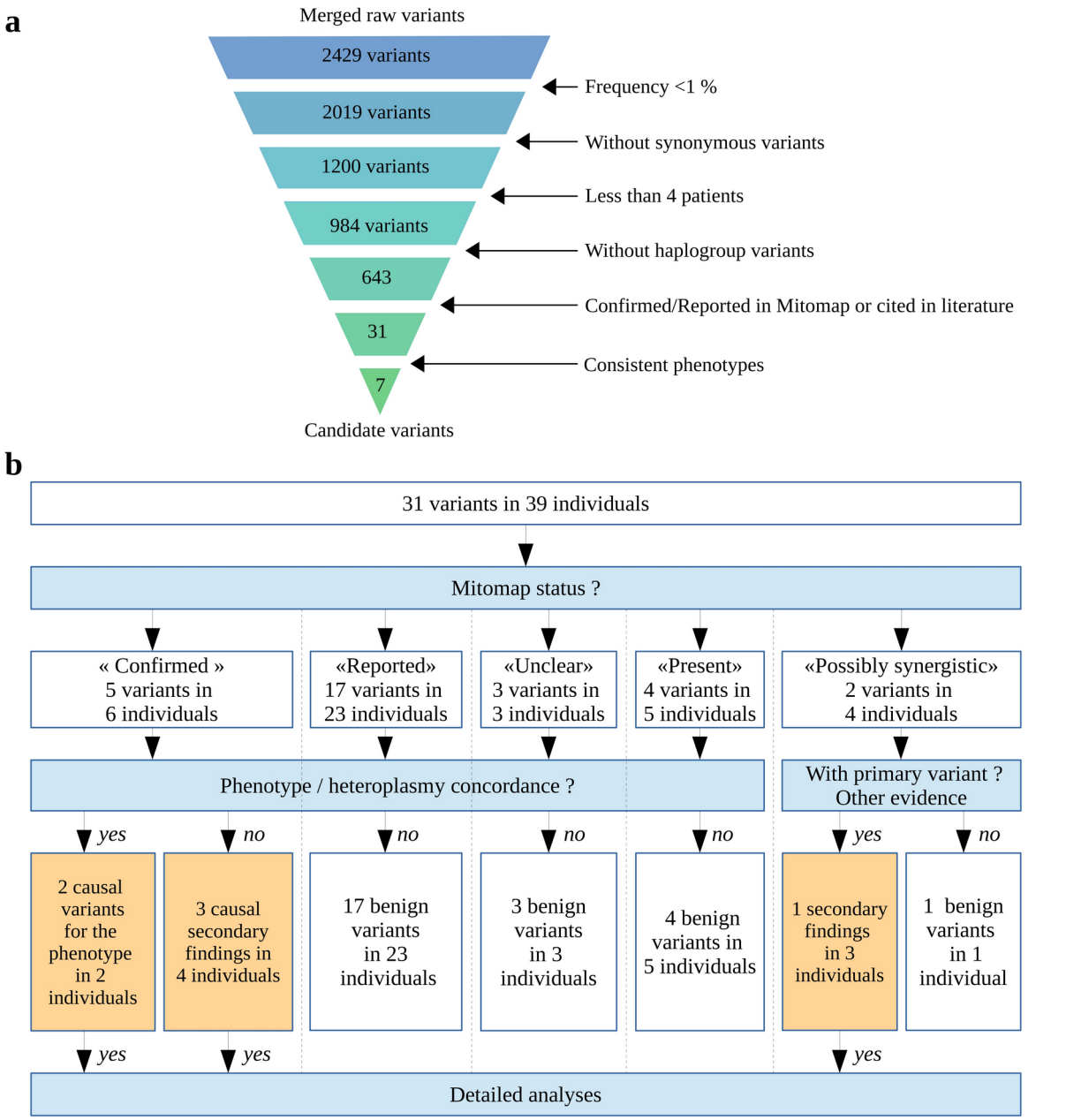


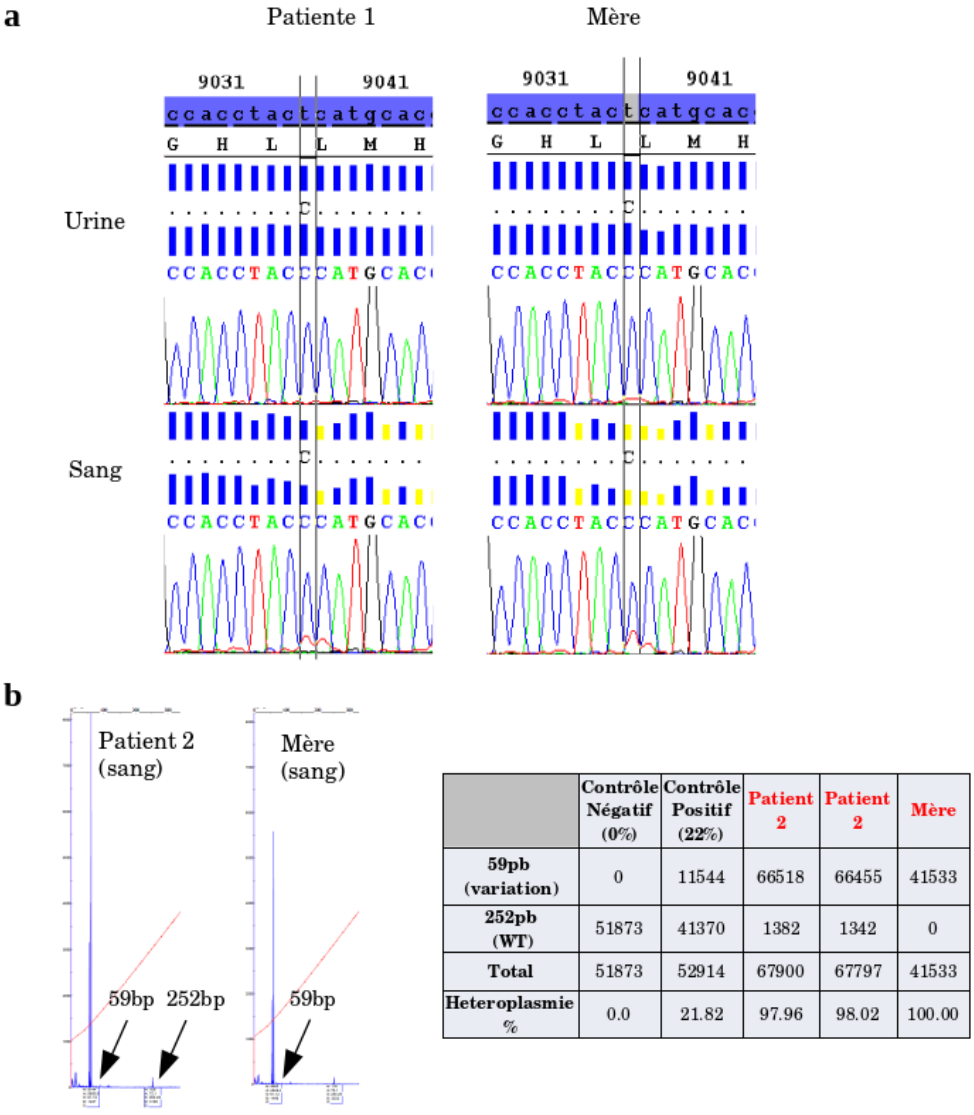
Figure 37 : Identification des variations mitochondriales candidates

a) Filtres appliqués aux variations mitochondriales identifiées chez les 928 patients de la cohorte. Les variations trop fréquentes (>1 % dans Mitomap et >4 fois dans la cohorte), les variations synonymes et les variations d'haplogroupes ont été éliminées informatiquement de la liste des candidats. Ainsi 73,5 % (n=1786) ont été filtrées sans étudier le phénotype des patients concernés. Puis les statuts Mitomap et la littérature ont été étudiés manuellement et comparés aux données des patients. 99,71 % des variations mitochondriales initiales ont ainsi été éliminées. b) Détails de l'analyse des variations classifiées dans Mitomap et retrouvée au sein de la cohorte. 31 variations ont été identifiées chez 39 patients après les filtres automatiques. La concordance phénotype et/ou taux d'hétéroplasmie entre les cas décrits et les patientes a ensuite été analysée. Les variations « probablement synergiques » n'ont pas été retenues en l'absence de leur variation primaire. A l'issue de l'ensemble des filtres, 7 variations identifiées chez 9 patients ont été retenues et vérifiées par une deuxième méthode moléculaire.

Les 5 variations « confirmées » NC_012920.1:m.1494C>T (*MT-RNR1*; OMIM 561000.0004), NC_012920.1:m.1555A>G (*MT-RNR1*; OMIM 561000.0001), NC_012920.1:m.9035T>C (*MT-ATP6*; OMIM 516060), NC_012920.1:m.11778G>A (*MT-ND4*; OMIM 516003.0001) et NC_012920.1:m.14484T>C (*MT-ND6*; OMIM 516006.0001) identifiées chez 6 patients, ont été considérées comme pathogènes, causale ou contributive, chez 2/6 patients et pathogènes secondaires chez 4/6 patients :

- La variation m.9035T>C a été confirmée par séquençage Sanger à l'état hétéroplasmique dans le sang et homoplasmique dans les urines (**Fig. 38a**) de la jeune femme de 30 ans atteinte de difficultés d'apprentissage, d'ataxie et de neuropathie axonale depuis 10 ans (**Tableau 7, patiente 1**). Les analyses effectuées *a posteriori* chez la mère ont donné les mêmes résultats chez cette femme de 59 ans diagnostiquée 6 ans plus tôt comme atteinte d'une neuropathie axonale isolée (F**ig. 38a**).

- La variation m.11778G>A a été confirmée par PCR-RFLP, à l'état hétéroplasmique dans le sang et homoplasmique dans les fibroblastes d'un patient de 30 ans (**Tableau 7, patient 2 et Fig. 38b**). Ce jeune homme appartient à la cohorte nouvellement décrite de patients atteints de dysplasie neuroectodermique en mosaïque avec des anomalies pigmentaires, dentaires, des extrémités et cérébrales, ainsi qu'une malformation oculaire et une perte de vision. Cette pathologie en mosaïque est causée par une variation post-zygotique dans le gène *RHOA* (Vabres et al., 2019). Néanmoins, ce patient était le seul de cette nouvelle cohorte à avoir une atteinte oculaire aussi prononcée et le seul porteur de cette variation mitochondriale. Il s'agit donc d'un exemple de double-hit, permettant d'expliquer un tableau clinique complexe. La PCR-RFLP effectuée sur le sang de sa mère a mis en évidence cette variation a l'état homoplasmique chez cette femme décrite comme asymptomatique (**Fig. 38b**).



	Contrôle Négatif (0%)	Contrôle Positif (22%)	Patient 2	Patient 2	Mère
59pb (variation)	0	11544	66518	66455	41533
252pb (WT)	51873	41370	1382	1342	0
Total	51873	52914	67900	67797	41533
Heteroplasmie %	0.0	21.82	97.96	98.02	100.00

Figure 38 : Validation par Sanger ou PCR-RFLP des variations m.9035T>C et m.11778G>A *a) Séquençage Sanger sur ADN extrait de sang et d'urine de la variation m.9035T>C chez la patiente 1 et sa mère. Leurs taux d'hétéroplasmie est proche de 100 %. b) PCR-RFLP pour valider la variation m.11778G>A sur ADN extrait de sang et de fibroblastes pour le patient 2 et de sang pour sa mère. Le taux d'hétéroplasmie est proche de 100 % pour le patient (pic à 59 pb et à 252 pb). La variation est homoplasmique pour la mère (pic à 59 pb uniquement).*

- Les 3 autres variations « confirmées » ont été considérées comme des données secondaires car les patients concernés ne présentaient pas de phénotypes similaires à ceux décrits dans la littérature (**Tableau 7, patients 3 à 6**). Les variations m.1494C>T et m.1555A>G sont décrites comme impliquées dans la surdité induite par les aminosides.

Elles ont été identifiées chez un fœtus de sexe masculin avec anomalies du développement (m.1494C>T à l’état homoplasmique), chez un patient de 19 ans présentant une ataxie (m.1555A>G à l’état hétéroplasmique) et un patient de 15 ans atteint d’un syndrome polymalformatif (m.1555A>G à l’état homoplasmique). Leurs histoires familiales respectives ne présentaient pas d’antécédent de surdité. La variation m.14484T>C a été décrite dans la neuropathie optique de Leber (LHON ; OMIM 535000). Elle a été identifiée à l’état homoplasmique chez un patient de 34 ans avec dystrophie musculaire sans atteinte oculaire.

Patient (sexe)	Age (ans)	Haplogroupe	Variation ADNmt	Statut Mitomap	Hétéroplasmie détectée par le pipeline (lectures mutées/ lectures totales)	Pathologies associées aux variations	Phénotype des patients
1 (F)	30	U5a1i1	m.9035T>C	Confirmed	100 % 17/17	Ataxie	Ataxie
2 (M)	30	H6a1b3	m.11778G>A	Confirmed	100 % 432/432	LHON	Atteinte oculaire + dysplasie neuro-ectodermique en mosaïque
3 (M)	Fœtus	H23	m.1494C>T	Confirmed	100 % 17/17	Surdité induite par les aminosides	Anomalies du développement
4 (M)	19	H1c	m.1555A>G	Confirmed	66,67 % 18/27	Surdité induite par les aminosides	Ataxie + nystagmus
5 (M)	15	H1be	m.1555A>G	Confirmed	100 % 15/15	Surdité induite par les aminosides	Syndrome polymalformatif
6 (M)	34	X1'2'3	m.14484T>C	Confirmed	100 % 5/5	LHON	Dystrophie musculaire
7 (F)	20	L2a1a2	m.14502T>C	Reported - possibly Synergistic	100 % 10/10	LHON	Syndrome de Rubinstein-Taybi
8 (F)	5	M10a1a1b1	m.14502T>C	Reported - possibly Synergistic	97,95 % 48/49	LHON	Syndrome de déficience en GLUT1
9 (F)	28	HV0a1	m.14502T>C	Reported - possibly Synergistic	100 % 3/3	LHON	Déficience intellectuelle + anomalies du développement

LHON: Leber's Hereditary Optic Neuropathy. La séquence de référence est NC_012920.1. Les haplogroupes ont été identifiés par HaploGrep 2.0

Tableau 7 : Variations pathogènes causales et secondaires identifiées au sein de la cohorte

Dix-sept variations « rapportées » dans Mitomap ont été identifiées chez 23 patients de la cohorte (**Fig. 37b**). Seize d'entre elles n'ont pas été retenues car le phénotype des patients ou

les taux d'hétéroplasmie n'étaient pas concordants avec ceux décrits dans la littérature. La variation NC_012920.1:m.5567T>C (*MT-TW*; OMIM 590095) a été mise en évidence à l'état homoplasmique chez 2 patients non apparentés, âgés de 55 et 14 ans, tous deux atteints d'une atrophie cérébelleuse accompagnée d'une ataxie et de même haplogroupe (K1a). La base de données Mitomap indique un score MitoTIP de 32,70 % soit une prédiction de caractère probablement bénin. De plus, cette variation était retrouvée chez 41 individus de la base de données GenBank. Enfin, le fait que les 2 patients partagent le même haplogroupe ne joue pas en faveur d'un caractère pathogène de cette variation, mais indiquerait plutôt une histoire évolutive commune. La variation m.5567T>C n'a donc finalement pas été retenue dans la liste des candidats.

Les variations des 11 derniers patients n'ont pas été retenues car le phénotype des patients ou les taux d'hétéroplasmie n'étaient pas concordants avec ceux décrits dans la littérature. Premièrement, trois variations retrouvées chez 3 patients avaient un statut « incertain » dans Mitomap ou présentaient des descriptions contradictoires. Deux patients porteurs de deux de ces variations n'ont pas été conservés dans la liste des candidats en raison de leurs phénotypes incompatibles avec ceux de la littérature. Le dernier individu était un patient atteint d'une surdité neurosensorielle causée par une dilatation des vestibules et aqueducs. Il est porteur de la variation homoplasmique NC_012920.1:m.8348A>G (*MT-TK*; OMIM 590060) pour laquelle le phénotype décrit est très différent de celui du patient. De plus la pathogénicité de cette variation demeure incertaine : elle n'a donc pas été retenue. Deuxièmement, deux variations identifiées chez 4 patients, NC_012920.1:m.11696G>A (*MT-ND4*; OMIM 516003) et NC_012920.1:m.14502T>C (*MT-ND6*; OMIM 516006), ont été décrites comme « possiblement synergiques » quand elles sont associées aux variations pathogènes primaires m.1555A>G (surdité induite par les aminosides), m.11778G>A (LHON) et m.14484T>C (LHON). Ces variations synergiques moduleraient le phénotype des patients induits par les variations primaires. Mais il existe également des cas de LHON directement provoqués par la variation homoplasmique m.14502T>C avec une pénétrance bien plus faible que celle observée pour les variations primaires de LHON (Zhao et al., 2009). La variation m.14502T>C a été mises en évidence chez 3 patients atteints respectivement des syndromes de

Rubinstein-Taybi, de déficience en GLUT1 et de déficience intellectuelle avec anomalies du développement. Elle a été considérée comme donnée secondaire et confirmée par une deuxième méthode moléculaire lorsque l'ADN était disponible. La variation m.11696G>A n'a pas été retenue dans la liste des candidats en raison de l'absence de variation primaire et d'un phénotype non concordant. Les 4 dernières variations identifiées chez 5 patients étaient présentes dans la base de données Mitomap mais n'étaient pas classifiées. Le phénotype des patients ne concordait pas avec ceux décrits dans la littérature : les variations n'ont donc pas été retenues.

Pour chaque variation candidate retenue, une étude des régions pseudomitochondriales correspondantes a été réalisée. Aucune variation nucléaire n'y a été identifiée. Les variations détectées par le nouveau pipeline d'analyse de données d'exome étaient bien des variations mitochondriales. Les validations moléculaires réalisées *a posteriori* l'ont d'ailleurs confirmé. L'analyse de la profondeur à chaque position de l'ADNmt a établi que l'intégralité de sa séquence était couverte par des données capturées indirectement. Enfin, l'étude des régions pseudomitochondriales a montré que l'intégralité de l'ADNmt peut être capturé par les kits de capture Agilent utilisés pour séquencer cette cohorte de 928 patients (**Fig. 15**).

A la suite de ce projet, le pipeline d'analyse de l'ADN mitochondrial à partir des données d'exome a été déployé en routine au laboratoire. Depuis, deux nouveaux cas positifs ont été identifiés :

- Le premier cas est une patiente de 10 ans atteinte d'atrophie optique bilatérale et porteuse de la variation NC_012920.1:m.13094T>C (*MT-ND5*; OMIM 516005). Cette dernière a été décrite dans des cas de LHON à l'état homoplasmique ou hétéroplasmique. Le séquençage Sanger a confirmé la présence de la variation à l'état hétéroplasmique dans les urines de la patiente (75 %) et de sa mère asymptomatique (25 %).
- Le deuxième cas est celui d'une patiente de 6 ans atteinte d'une hypoplasie du nerf optique et issue d'une union consanguine. Elle est porteuse d'une variation m.11778G>A identifiée par le pipeline bioinformatique à l'état homoplasmique. La validation par PCR-RFLP a confirmé la présence de cette variation à l'état

homoplasmique chez la patiente et à l'état hétéroplasmique (77 %) chez la mère asymptomatique.

L'ensemble de ces travaux sur l'ADNmt a fait l'objet d'une publication dans *Human Mutation* (**Article 2**).

IV- DISCUSSION : L'ANALYSE DE L'ADNMT, UNE LIMITE BIOINFORMATIQUE REPOUSSÉE

L'analyse de données de séquençage indirect de l'ADNmt au sein de la cohorte de 928 patients avec DI et/ou AD a permis d'identifier 2 variations pathogènes, causale ou contributive, chez 2 patients (m.9035T>C and m.11778G>A). Quatre variations secondaires ont également été mises en évidence chez 7 patients. Suite à cette première analyse rétrospective, ce pipeline a également permis d'identifier 2 cas positifs supplémentaires de manière prospective. Durant cette étude des défis biologiques et bioinformatiques ont dû être relevés. En effet, bien qu'extraits et séquencés simultanément (Samuels et al., 2013), les ADN nucléaire et mitochondrial ne peuvent pas être analysés par le même processus en raison de leurs spécificités. Tout d'abord le pipeline bioinformatique a été adapté pour travailler sur des données de séquençage indirect de l'ADNmt avec une modification de la référence, un changement de bases de données, et l'ajout d'étapes spécifiques à l'étude de l'ADNmt. Ces dernières concernent l'extraction du génome mitochondrial des données d'alignement, la détermination de l'haplogroupe de chaque individu étudié et le calcul du taux d'hétéroplasmie de chaque variation identifiée.

En premier lieu, le génome de référence a été modifié. En effet, la majorité des laboratoires continuent d'utiliser la référence du génome humain GRCh37/hg19. Mais seule la version GRCh38 contient l'actuelle séquence de référence de l'ADNmt (Ye et al., 2014). L'alignement sur le génome complet est considéré comme la meilleure approche pour réduire le sur-

alignement des régions pseudomitochondriales à l'origine de biais dans la détermination du taux d'hétéroplasmie. Il a d'ailleurs fallu vérifier que les séquences mitochondriales détectées n'étaient liées pas aux régions pseudomitochondriales, en raison de la capture indirecte. Les résultats des contrôles positifs ont en effet confirmé que les variations mitochondriales causales identifiées n'étaient pas d'origine pseudo-mitochondriale. De plus, les positions pseudomitochondriales équivalentes à celles des variations mitochondriales causales identifiées au sein de la cohorte des 928 patients ont été vérifiées manuellement. Ainsi, les variations détectées étaient bien localisées sur le génome mitochondrial.

Par ailleurs, l'analyse des fichiers VCF obtenus par NGSmt pour les contrôles 3 et 4 et les patients 1, 7 et 9 ont montré que le « variant caller » HaplotypeCaller, bien que non spécifique de la mitochondrie, permettait de mettre en évidence les variations d'intérêt au sein du génome mitochondrial (données non présentées).

Une des adaptations les plus importantes pour pouvoir réaliser l'étude de patients, sans atteinte mitochondriale suspectée, est la détermination de l'haplogroupe de chaque patient. Cette étape permet en premier lieu de vérifier la qualité de la séquence d'ADNmt reconstruite (Diroma et al., 2014). Mais elle offre également la possibilité d'ajouter un filtre pour prioriser l'interprétation des variations mitochondriales. En effet, ces variations dépendent de l'origine du patient et ne sont donc pas prises en compte lors de la suite de l'analyse. Cette méthode a déjà été décrite (Santorsola et al., 2016) lors de l'étude de cellules tumorales (Calabrese et al., 2016) ou d'une cohorte d'individus suspectés d'être atteints d'une pathologie mitochondriale (Patowary et al., 2017).

La comparaison, en oncologie, de variations de l'ADNmt germinales ou somatiques souligne la nécessité de filtres afin de prioriser l'analyse de ces variations. Peuvent être cités en exemple les filtres sur la fréquence ou sur l'haplogroupe (Calabrese et al., 2016) que l'on peut également retrouver lors de l'étude d'une cohorte d'individus suspectés d'être atteints d'une pathologie mitochondriale. Filtrer les variations d'haplogroupes a permis de prioriser les variations de l'ADNmt, de confirmer la causalité de variations connues dans une pathologie de type LHON (Santorsola et al., 2016) ou de mettre en évidence des variations candidates pour l'autisme (Patowary et al., 2017), pathologie qui serait liée à une anomalie de fonctionnement de la

mitochondrie. Lors de l'analyse d'une cohorte de patients sans pathologie mitochondriale suspectée, comme cela a été réalisée durant cette étude, le filtre des variations d'haplogroupes permet d'éliminer des polymorphismes non filtrés précédemment. En effet, certaines variations d'haplogroupes sont peu fréquentes dans la population générale.

En dehors de l'analyse bioinformatique, l'étude a présenté un défi biologique lors de l'interprétation des variations mitochondriales identifiées. L'analyse des 5 variations au statut « confirmé » dans Mitomap a été facilitée par les nombreuses études démontrant leur pathogénicité. Lorsque la deuxième méthode de validation moléculaire a confirmé la présence de ces variations à des taux d'hétéroplasmie attendus et que les phénotypes des patients concordaient avec ceux décrit, ces variations ont été considérée comme causales (**Tableau 7 Patients 1 et 2**). Mais lorsque les phénotypes ne concordaient pas avec l'indication de l'analyse de l'exome, les variations ont été classées comme données secondaires (**Tableau 7 Patients 3 à 9**), car il existe une pénétrance incomplète des variations d'ADNmt. Ainsi cette étude a permis d'identifier chez la patiente 1 et sa mère la variation homoplasmique m.9035T>C responsable de son ataxie et du phénotype plus modéré décrit chez sa mère (**Fig. 38a**). En effet cette variation, au statut « confirmé » a été décrite dans l'ataxie (Pfeffer et al., 2012). Par ailleurs, ces travaux ont permis d'identifier la variation homoplasmique m.11778G>A chez le patient 2 atteint et sa mère asymptomatique (**Fig. 38b**). Cette variation a été décrite dans l'atrophie optique de Leber à faible pénétrance (Wallace et al., 1988). Ce patient est atteint d'une pathologie neuroectodermique en mosaïque causée par une variation post-zygotique hétérozygote dans le gène *RHOA* (Vabres et al., 2019). L'identification de ce double-hit a permis de « préciser » ce syndrome, plus particulièrement les atteintes oculaires, apportant des informations importantes pour le conseil génétique.

Les autres variations identifiées au cours de cette étude avaient un statut « rapporté », « probablement synergique », « uncertain » dans Mitomap ou n'étaient pas classifiées. Leur pathogénicité chez les patients concernés n'est pas clairement établie en raison de scores de prédiction faibles, d'absence de concordance phénotypique ou une absence de variation primaire.

La pénétrance incomplète constitue un des défis biologiques dans l'interprétation des variations mitochondriales. L'expression d'une variation nucléaire ou mitochondriale est dépendante, entre autres, de son environnement génétique (Cooper et al., 2013). Ainsi, la

position de la variation au sein du gène ou de la protéine peut modifier son impact. La pathogénicité d’une variation peut également être renforcée par la présence d’une autre variation au sein du même gène. De même, il existe des variations qui peuvent contrer les effets d’une variation pathogène. Des gènes modificateurs peuvent aussi participer à la modification de l’expression d’un gène et des variations pathogènes qu’il contient. Il existe également des pathologies à hérédité oligogénique pour lesquelles une seule variation pathogène ne permet pas l’expression de la maladie. L’implication d’une nouvelle variation de l’ADNmt, non précédemment décrite, peut-être difficile à retenir, car les études de ségrégation familiale peuvent donner des informations limitées compte tenu des possibles pénétrance incomplète et expressivité variable.

Un autre point important lors de l’étude des 928 patients, sans atteinte mitochondriale suspectée, est le taux d’hétéroplasmie. Celui-ci dépend du choix du tissu prélevé. La séquence de l’ADNmt étudiée lors de ces travaux provient de données indirectes d’exome. Les régions du génome mitochondrial n’ont donc pas été couvertes de manière homogène. La couverture moyenne de l’ADNmt était d’ailleurs de 50X, valeur inférieure à la moyenne de 100X obtenue pour l’exome nucléaire. Cette différence peut être due au choix des kits de capture car ils présentent des designs et des séquences cibles différentes conduisant à une variation des séquences off-target séquencées et donc à une couverture hétérogène de l’ADNmt. La profondeur et la couverture obtenues par cette méthode n’étant pas optimales, comparée à une méthode directe d’analyse de l’ADNmt, certaines variations peuvent ne pas être détectées ou des taux d’hétéroplasmie peuvent être biaisés. En effet, le séquençage ciblé de l’ADNmt (NGSmt) est une méthode spécifique pour ce type d’analyse par la capture et l’amplification directes de cet ADN. La détermination du taux d’hétéroplasmie est donc plus précise. Le NGSmt permet également, en augmentant la profondeur, de déterminer des taux d’hétéroplasmie très faibles proches du bruit de fond (~1 %). Il est donc conseillé d’utiliser une méthode directe d’analyse de l’ADNmt pour confirmer le résultat négatif des 919 patients restants de notre étude. Mais l’analyse des 4 contrôles positifs démontre que cette méthode indirecte peut identifier le statut hétéroplasmique ou homoplasmique ainsi que le taux d’hétéroplasmie d’une variation mitochondriale même en cas de faible profondeur. En effet le

coefficient de corrélation de Pearson entre les taux obtenus par cette méthode et ceux déterminés par une méthode ciblée est de 0,943. De plus, la comparaison des variations détectées (position + variation) par la méthode indirecte et par le NGSmt montre une bonne corrélation des résultats. Enfin cette méthode indirecte a permis une augmentation du taux diagnostique de l’exome car de nouveaux cas ont été résolus. Une analyse des données de génome par cette méthode pourrait obtenir de meilleures couverture et profondeur. En effet il s’agit d’une méthode de séquençage sans étape d’amplification PCR, c’est-à-dire sans enrichissement spécifique de régions nucléaires. Toutes les molécules d’ADN au sein de la cellule, dont l’ADNmt, sont capturées. Comme il y a 10 à 100 fois plus de molécules d’ADNmt que d’ADN nucléaire dans chaque cellule (Dinwiddie et al., 2013) la couverture et la profondeur seront améliorées par cette méthode. La détection des variations, étape sensible à la profondeur, et la détermination de leur taux d'hétéroplasmie seront donc plus précises.
La précision du calcul du taux d’hétéroplasmie dépend de la profondeur de séquençage c’est-à-dire du nombre de lectures alignées à la position considérée (Griffin et al., 2014). De plus ce taux varie en fonction du tissu prélevé. Pour cette étude l’ADN a été extrait en grande majorité du sang au sein duquel les variations mitochondriales pourraient être absentes ou avec un taux d’hétéroplasmie trop faible pour être détectées. L’analyse de l’ADNmt par méthode directe reste donc conseillée chez les individus suspectés d’atteintes mitochondriales.

Le nombre de variations secondaires identifiées lors de cette étude est supérieur à celui des diagnostics posés. Ces variations sont décrites comme responsables de pathologies à pénétrance incomplète de 0,53 % pour le LHON contre 0,29 % dans la population ; et de 0,3 %, proche de la valeur de 0,28 % dans la population, pour la surdité induite par les aminosides. La question du rendu de ces données secondaires reste ouverte notamment en raison des conseils de prévention qui pourraient en découler, à contrebalancer avec l’inquiétude qui peut être générée par ces pathologies qui pourraient ne jamais se déclarer.

En 2017, Bergant et al. et Patowary et al . ont étudié l’ADNmt à partir de données d’exome. Dans la première étude, les auteurs cherchaient à augmenter le taux diagnostique de l’exome en exploitant les données non traitées (ADNmt, CNV ; etc.) lors d’une analyse

d’exome nucléaire (Bergant et al., 2018). 1059 patients ont ainsi été analysés. Si la cohorte comportait plusieurs types de pathologies, les atteintes neurologiques représentaient 39 % des patients. Le principe de cette analyse du génome mitochondrial consistait en la reconstitution de la séquence de l’ADNmt du patient avant son analyse par l’outil MITOMASTER accessible sur le site de Mitomap. Les variations ont alors été annotées avec les informations de la base de données Mitomap. Seules les variations dont la fréquence était inférieure à 1 % dans leur base de données internes (population slovène) et dans GenBank, ainsi que les celles d’une profondeur d’au moins 10 reads ont été conservées. Les auteurs se sont concentrés sur les variations déjà connues en pathologies mitochondriales (de statut « confirmé » dans Mitomap). L’outil MITOMASTER détermine également l’haplogroupe mais aucun filtre sur les variations impliquées dans ce dernier n’a été décrit dans cette étude. Au cours de cette étude, 3 variations causales ont été identifiées soit une augmentation diagnostique de 0,28 %. Malgré les différences entre nos deux approches, les taux diagnostiques sont similaires. Il s’agit en effet de deux cohortes de patients non spécifiquement atteints de pathologies mitochondriales analysées par une méthode indirecte. Dans la seconde étude, les auteurs ont étudié 10 familles multiplex atteintes d’autisme pour étudier l’implication de la mitochondrie dans les TSA (Patowary et al., 2017). Ils ont tout d’abord reconfirmé que l’ADNmt pouvait être étudié à partir de données d’exome et ont décrit les défis rencontrés : hétérogénéité phénotypique, âge d’apparition des pathologies, pénétrance incomplète ou taux d’hétéroplasmie. Ils ont choisi l’outil MtoolBox pour effectuer leur analyse qui leur a permis d’identifier 5 variations d’intérêt potentiellement impliquées dans l’autisme. Il s’agit d’une méthode différente d’analyse de l’ADNmt. Contrairement à notre méthode, l’alignement des données de séquençage d’exome est réalisé d’abord sur la référence mitochondriale puis sur hg19. Seules les lectures alignées de façon unique sur le génome mitochondrial de référence sont conservées pour générer un fichier VCF, calculer les taux d’hétéroplasmie et reconstruire l’ADNmt complet. L’haplogroupe de chaque individu est alors identifié en alignant cette séquence sur les séquences consensus des macro-haplogroupes. Notre méthode ne reconstruit pas la séquence intégrale de l’ADNmt et détermine les haplogroupes grâce à l’outil HaploGrep2. Enfin les données de fréquence auxquelles ils ont fait appel proviennent de la base de données 1000 Génomes et non de Mitomap, qui est spécialisée dans la mitochondrie et qui contient plus d’individus séquencés. Le coefficient de corrélation entre MtoolBox et les méthodes ciblées (r=0,47) est plus faible qu’entre ces

dernières et notre méthode (r=0,91). Notre pipeline d'analyse de l'ADNmt à partir de données d'exome est donc plus sensible.

Lors des révisions de notre article une nouvelle étude est parue. Il s'agit de l'analyse d'une cohorte de 2111 patients suspectés d'être atteints d'une pathologie mitochondriale (Wagner et al., 2019). La méthode a d'abord consisté en l'alignement des données d'exome sur la référence hg19 où la séquence de l'ADNmt a été remplacée par la référence rCRS. L'annotation de l'ADNmt a ensuite été réalisée par des scripts internes. Seules les variations au statut « confirmé » dans Mitomap ont été analysées. Le génome mitochondrial de 49 individus de la cohorte a également été séquencé de façon ciblée par PCR à longs fragments afin de permettre d'estimer la précision de l'appel de variants et de la détermination du taux d'hétéroplasmie. La profondeur moyenne du génome mitochondrial était de 58 ± 38X. Cette valeur est proche de celle obtenue avec notre méthode. La couverture était également hétérogène. L'étude portait d'une part sur le génome nucléaire et d'autre part sur l'ADNmt. L'analyse du premier a permis d'identifier 889/2111 (42,1 %) de diagnostics positifs et celle du second 38/2111 (1,8 %). Ce taux diagnostique plus élevé que pour notre étude ou celle de Bergant peut s'expliquer par la composition de la cohorte. Le regroupement de patients spécifiquement atteints d'une pathologie mitochondriale a en effet « enrichi » les résultats de variations mitochondriales causales. Ils ont néanmoins rappelé qu'un résultat négatif de cette méthode d'analyse indirecte ne pouvait pas exclure la présence d'une variation mitochondriale causale. Ils recommandaient donc des analyses plus ciblées en cas de suspicion de pathologie mitochondriale.

L'application de cette méthode aux données de génome nécessitera une mise au point bioinformatique notamment pour adapter ce pipeline à l'analyse d'un volume de données plus important (mémoire, temps, etc.).

Ces travaux ont donc permis de développer un pipeline d'analyse de l'ADNmt à partir de données préexistantes de séquençage d'exome. Cette méthode permet de repousser une des limites bioinformatiques à l'origine de l'errance diagnostique (Thevenon et al., 2016) et d'augmenter ainsi le nombre de cas résolus (0,2 %). Mais il reste des informations

supplémentaires non extraites des données d’exome et qui peuvent elles aussi participer à l’amélioration du taux diagnostique. La dernière partie du travail de thèse a donc consisté à repousser une deuxième limite bioinformatique qui est la détection des éléments mobiles à partir des données de ES.

Article 2

Deciphering exome sequencing data: bringing mitochondrial DNA variants to light.

Hum Mutat, Dec 2019;40(12), 2430-2443

Philippine Garret, Celine Bris, Vincent Procaccio, Patrizia Amati Bonneau, Pierre Vabres, Nada Houcinat, Emilie Tisserant, François Feillet, Ange-Line Bruel, Virginie Quéré, Christophe Philippe, Arthur Sorlin, Frédéric Tran MauThem, Antonio Vitobello, JeanMarc Costa, Aïcha Boughalem, Detlef Trost, Laurence Faivre, Christel ThauvinRobinet, Yannis Duffourd.

Received: 5 January 2019 | Revised: 27 June 2019 | Accepted: 30 July 2019

DOI: 10.1002/humu.23885

METHODS

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Deciphering exome sequencing data: Bringing mitochondrial DNA variants to light

Philippine Garret[1,2,3] | Céline Bris[4,5] | Vincent Procaccio[4,5] | Patrizia Amati-Bonneau[4,5] | Pierre Vabres[1,6,7] | Nada Houcinat[1,2,8,9] | Emilie Tisserant[1,2] | François Feillet[10,11,12] | Ange-Line Bruel[1,2] | Virginie Quéré[1,6] | Christophe Philippe[1,2] | Arthur Sorlin[1,2,6,8] | Frédéric Tran Mau-Them[1,2,8,9] | Antonio Vitobello[1,2] | Jean-Marc Costa[3] | Aïcha Boughalem[3] | Detlef Trost[3] | Laurence Faivre[1,8,13] | Christel Thauvin-Robinet[1,2,9*] | Yannis Duffourd[1,2*]

[1]INSERM–University of Burgundy-Franche Comté, UMR1231 GAD, Dijon, France

[2]Unité Fonctionnelle Innovation en Diagnostic génomique des maladies rares, FHU-TRANSLAD, Dijon University Hospital, Dijon, France

[3]Laboratoire CERBA, Saint-Ouen-l'Aumône, France

[4]Institut MITOVASC, UMR CNRS 6015-INSERM1083, University of Angers, Angers, France

[5]Centre de Référence maladies mitochondriales, CHU Angers, Angers, France

[6]Centre de Référence Maladies Rares « Maladies Dermatologiques en Mosaïque », Service de dermatologie, FHU-TRANSLAD, Dijon University Hospital, Dijon, France

[7]Service Dermatologie, CHU Dijon Bourgogne, Dijon, France

[8]Centre de Référence Maladies Rares « Anomalies du développement et syndromes malformatifs », Centre de Génétique, FHU-TRANSLAD, Dijon University Hospital, Dijon, France

[9]Centre de Référence Maladies Rares « déficience intellectuelle », Centre de Génétique, FHU-TRANSLAD, Dijon University Hospital, Dijon, France

[10]Service de Pédiatrie, Hôpital d'Enfants Brabois, CHRU Nancy, Vandoeuvre les Nancy, France

[11]INSERM–University of Lorraine–CHRU Nancy, UMRS 1256 NGERE, Nancy, France

[12]Centre de Références des maladies héréditaires du métabolisme, CHRU de Nancy, Nancy, France

[13]Centre de compétences des maladies mitochondriales, Dijon University Hospital, Dijon, France

Correspondence
Christel Thauvin-Robinet and Yannis Duffourd, Inserm UMR1231 GAD–Genetics of Developmental Disorders, UFR Sciences de Santé–Bâtiment B3, 15 Boulevard Maréchal de Lattre de Tassigny, 21070 Dijon Cedex, France.
Email: christel.thauvin@chu-dijon.fr (C.T.-R.) and yannis.duffourd@u-bourgogne.fr (Y.D.)

Funding information
CIFRE (ANRT); PARI 2017; Conseil Régional de Bourgogne; FEDER 2017; Laboratoire Cerba

Abstract

The expanding use of exome sequencing (ES) in diagnosis generates a huge amount of data, including untargeted mitochondrial DNA (mtDNA) sequences. We developed a strategy to deeply study ES data, focusing on the mtDNA genome on a large unspecific cohort to increase diagnostic yield. A targeted bioinformatics pipeline assembled mitochondrial genome from ES data to detect pathogenic mtDNA variants in parallel with the "in-house" nuclear exome pipeline. mtDNA data coming from off-target sequences (indirect sequencing) were extracted from the BAM files in 928 individuals with developmental and/or neurological anomalies. The mtDNA variants were filtered out based on database information, cohort frequencies, haplogroups and protein consequences. Two homoplasmic pathogenic variants (m.9035T>C and m.11778G>A) were identified in 2 out of 928 unrelated individuals (0.2%): the

*Christel Thauvin-Robinet and Yannis Duffourd contributed equally to this work.

m.9035T>C (*MT-ATP6*) variant in a female with ataxia and the m.11778G>A (*MT-ND4*) variant in a male with a complex mosaic disorder and a severe ophthalmological phenotype, uncovering undiagnosed Leber's hereditary optic neuropathy (LHON). Seven secondary findings were also found, predisposing to deafness or LHON, in 7 out of 928 individuals (0.75%). This study demonstrates the usefulness of including a targeted strategy in ES pipeline to detect mtDNA variants, improving results in diagnosis and research, without resampling patients and performing targeted mtDNA strategies.

KEYWORDS
bioinformatics, ES data, mtDNA, pipeline

1 | INTRODUCTION

Mitochondrial disorders make up a vast group of diverse pathologies affecting 1 out of 5,000 live births (Bannwarth et al., 2013). Since mitochondrial organelles are present in every human cell, all organs can be concerned by mitochondrial dysfunction. The clinical spectrum is highly heterogeneous, varying from one to several symptoms including deafness, retinopathy, diabetes, myopathy, epilepsy, and renal, cardiac, or hepatic dysfunction. Metabolic investigation can reveal blood and/or cerebrospinal fluid (CSF) lactate increase, and muscle biopsy can evidence respiratory chain anomalies and/or ragged-red fibers. Mitochondrial disorders are particular that they can be transmitted by either Mendelian or mitochondrial inheritance since mitochondria are complex cellular organelles composed of multiple mitochondrial proteins encoded either by nuclear or mitochondrial DNA (mtDNA genes; Dinwiddie et al., 2013).

To date, more than 300 nuclear genes have been implicated in mitochondrial disorders (Gorman et al., 2016), mainly with the autosomal recessive mode of inheritance (Chinnery, 2002). For the maternally inherited mtDNA, 37 genes are known to be involved in human diseases with large clinical and genetic variability (Dinwiddie et al., 2013). Indeed, different phenotypes are linked to the same mtDNA variant and the same phenotype to different mtDNA variants. For example, the GenBank NC_012920.1:m.3243A>G variant (*MT-TL1;* MIM# 590050.0001; dbSNP Build 152: rs199474657 [Sherry et al., (2001)]) causes different phenotypes including mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS; MIM# 540000), chronic progressive external ophthalmoplegia (CPEO; MIM# 530000), and maternally inherited diabetes-deafness syndrome (MIDD; MIM# 520000). Conversely, MELAS syndrome may be the consequence of other pathogenic variants in *MT-TL1 (*e.g., NC_012920.1:m.3256C>T, NC_012920.1:m.3260A>G, NC_012920.1:m.3271T>C, and NC_012920.1:m.3291T>C; MIM# 590050.0003, 590050.0007, 590050.0002, and 590050; dbSNP Build 152: rs199474659, rs199474663, rs199474658, and rs869312463) or in other transfer ribonucleic acid (tRNA) genes (e.g., NC_012920.1:m.583G>A in *MT-TF;* MIM# 590070.0001; dbSNP Build 152: rs118203885 or NC_012920.1:m.12147G>A in *MT-TH*; MIM# 590040.0003; dbSNP Build 152: rs121434474). Individual clinical phenotypes also depend on mitochondrial specificities named heteroplasmy/homoplasmy (Bai & Wong, 2005). In a single individual, mtDNA sequence can be variable with some mtDNA variants present in mixed proportions with wild-type mtDNA, within a cell or a tissue: this condition is called heteroplasmy. Homoplasmy is when the mtDNA variant is fully present in all the cells.

In a diagnostic clinical setting, the expanding implementation of next-generation sequencing (NGS) has considerably improved the diagnosis of mitochondrial diseases over the last few years. Targeted whole-mtDNA sequencing initially allowed researchers to detect mtDNA variants easily (Y.He et al., 2010; Vasta, Ng, Turner, Shendure, & Hahn, 2009). Targeted strategies are also routinely used to detect mtDNA variants when a mitochondrial disorder has been clinically suspected, focusing on expected single-nucleotide variants (SNVs) or structural variants. Older technologies such as polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) currently detect single mtDNA variants (Holt, Harding, Petty, & Morgan-Hughes, 1990), which is useful for confirming low heteroplasmic rate variants (Vasta et al., 2009). For deletions and duplications, Southern Blot, real-time PCR, and long-range PCR are still commonly used (Colter-Mackie, Applegarth, Toone, & Gagnier, 1998; L.He et al., 2002).

The enrichment of exome sequencing (ES) capture was also proposed as a way to efficiently and specifically target all nuclear mitochondrial and mtDNA genes enriched fragments (Falk et al., 2012). Exome capture kits mainly used in a diagnostic scoop do not include mtDNA. It is possible to design exome capture kits for simultaneous nuclear exome and mitochondrial sequencing but they are rarely available in catalogs. Moreover, mtDNA can be sequenced indirectly, and therefore the mtDNA sequence can be extracted and reassembled from ES data (Samuels et al., 2013). This indirect mtDNA sequencing was initially used in cancer studies (Guo, Li, Li, Shyr, & Samuels, 2013; Zhang et al., 2016), with a level of efficiency similar to direct mtDNA sequencing in five breast cancer cell lines (Zhang et al., 2016). Single-nucleotide polymorphism and mutant loads were detected and quantified by indirect mtDNA sequencing, and the results obtained were almost the same as the direct

approach. In mitochondrial disorders, indirect mtDNA sequencing has only been reported in a few studies. For instance, Dinwiddie et al. (2013) reported the cases of three affected individuals highly suspected of having a mitochondrial disorder, and were able to identify mtDNA variants in two of them: the pathogenic homoplasmic (100%) NC_012920.1:m.8993T>C (p.L156P; *MT-ATP6;* MIM# 516060.0002; dbSNP Build 152: rs199476133) variant involved in Leigh syndrome and a variant of uncertain significance (VUS) novel heteroplasmic (25%) NC_012920.1:m.3754C>A (*MT-ND1*; MIM# 516000) variant. Increasing use of NGS requires the implementation of bioinformatics tools to process and analyze the huge amount of generated data. While the pipeline for nuclear variants identification from ES or genome sequencing (GS) data follows the known scheme described and recommended in GATK Best Practices (Van der Auwera et al., 2013), the indirect mtDNA sequencing and analysis based on ES data was not mentioned and presents some bioinformatics and biological limits. mtDNA can be studied with the same nuclear pipeline (Dinwiddie et al., 2013) from quality assessment to variant calling steps, but specific databases and processes are required for variant annotation. Moreover, over the course of human evolution, mtDNA acquired specific sets of associated variants defining mitochondrial haplotypes (or haplogroups) depending on geographical origins. These haplogroups are hierarchically organized in a tree comprising haplogroups and subhaplogroup branches. The effects of mtDNA variants may vary according to the mitochondrial genetic background (Wallace, Fan, & Procaccio, 2010). Thus, mitochondrial haplogroups need to be determined to properly interpret mitochondrial variants that notably differ from current nuclear variant interpretation. Haplogroup polymorphic variants would then be filtered out from the candidate variant list. Moreover, after mitochondrial endosymbiosis, mtDNA sequences have colonized the nuclear DNA (Calabrese, Simone, & Attimonelli, 2012). In primates, this can occur in several copies along the nuclear genome: these regions are nuclear mitochondrial DNA sequences (NUMTs) and their variants need to be filtered out to avoid false-positive mtDNA variant detection.

Different tools such as MitoSeek (Guo et al., 2013), mit-o-matic (Vellarikkal et al., 2015), or MtoolBox (Calabrese et al., 2014) have already been developed. While MitoSeek works only on remaining reads not aligned after nuclear genome analysis and does not identify haplogroups, mit-o-matic, and MtoolBox assign haplogroups. However, mit-o-matic aligns reads on revised Cambridge Reference Sequence (rCRS; GenBank NC_012920.1) human mtDNA reference only, and MtoolBox aligns sequences on the mitochondrial and nuclear genome separately. These tools have previously been used for indirect mtDNA sequencing data analysis, also with filtering of haplogroup variants, but only on a limited number of individuals suspected of autism (Patowary, Nesbitt, Archer, Bernier, & Brkanac, 2017).

Here, we present the creation and use of a bioinformatics pipeline with Burrows—Wheeler Aligner (BWA; Li & Durbin, 2009), SAMtools (Li et al., 2009), and genome analysis toolkit (GATK; McKenna et al., 2010) tools, using haplogroup determination to analyze the indirect mtDNA sequencing results issued from ES data in a large unspecific cohort of individuals with developmental anomalies (DA) and/or neurological disorders to identify whether mtDNA variants related to mitochondrial disorders can be found in individuals without identified nuclear variants.

2 | INDIVIDUALS AND METHODS

2.1 | Individuals

We gathered a cohort of 928 unrelated individuals with DA (81%) or primary neurological disorder (19%; 536 males and 392 females), mainly Caucasian and with ages ranging from unborn to 89 years (median 11 years), for which ES was performed in a diagnostic or research setting. ES was performed as previously described (Thevenon et al., 2016). The nuclear ES analysis underwent a positive result in 249/928 (26.8%). The mtDNA analysis was performed in the ES data of the 928 probands whatever the results of the nuclear ES analysis. When available, DNA from family members was used for segregation analysis. Informed written consent was obtained from individuals or parents for ES analysis.

2.2 | Positive controls

To validate this new mitochondrial bioinformatics pipeline, we analyzed ES data from four positive controls issued from DNA samples with a well-known mitochondrial variant previously identified by targeted sequencing (Table S1). For these samples, ES was performed with the Agilent CRE V2 or Agilent V5 enrichment kits, which are two of the different kits used in our cohort.

2.3 | Indirect mtDNA sequencing from ES data analysis

2.3.1 | Sequences alignment, file conversion, and mitochondrial genome extraction

We developed a bioinformatics pipeline devoted to mtDNA analysis of ES data. It was designed using the guidelines outlined in GATK Best Practices (Van der Auwera et al., 2013) and based on the "in-house" pipeline already employed for nuclear ES in our laboratory (Thevenon et al., 2016; Figure 1a). After quality control, pairs of fastq files were aligned on the GRCh38 reference because it contains the currently used mtDNA reference GenBank NC_012920.1 (Andrews et al., 1999), unlike the hg19 genome. Using the BWA-MEM (v.0.7.15; Li & Durbin, 2009), we chose the most recommended approach (Ye, Samuels, Clark, & Guo, 2014): complete alignment on GRCh38 reference.

After the alignment step, we used SAMtools (v.1.2; Li et al., 2009) to convert SAM into a sorted BAM, extracted the mitochondrial genome data and indexed it. Picard tools (v.2.4.1; Broad Institute, n.d) were then used to clean and mark optical duplicate reads on the extracted mtDNA data. Thereafter, base quality score recalibration was performed by the GATK (v.3.7; McKenna et al., 2010).

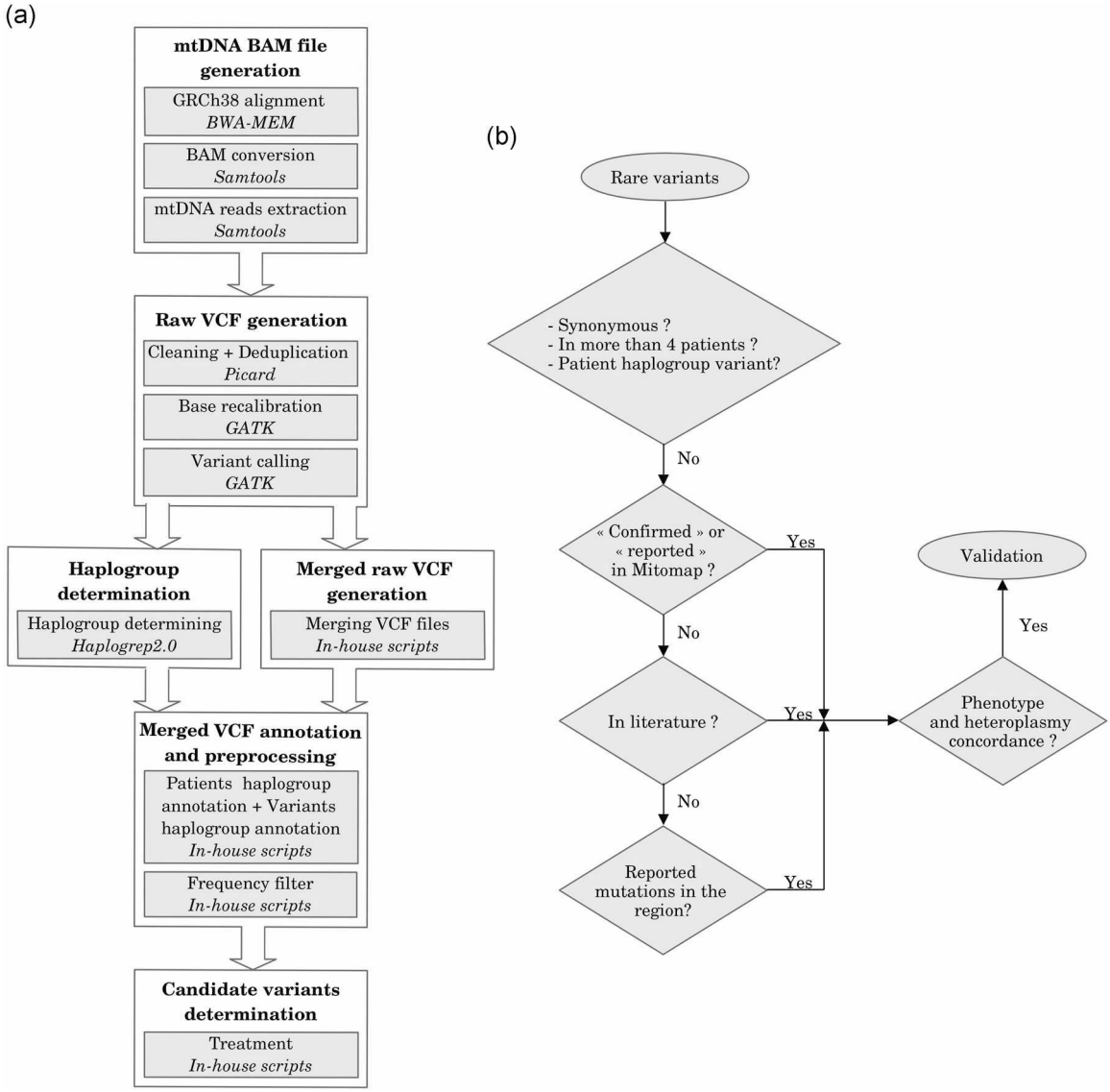


FIGURE 1 mtDNA analysis. (a) Schematic overview of the mtDNA pipeline. BAM files and raw VCF files generation scripts were based on our nDNA pipeline and the steps were completely automated (Thevenon et al., 2016). Merged VCF file generation required manual intervention due to the online HaploGrep 2.0 step. Only one raw merged file was needed. Variant frequency data were extracted and calculated from the Genbank sequence depository of Mitomap. In-house scripts allowed for a rapid filter of variants. Tools are noted in italics. (b) mtDNA rare variant analysis strategies. Variants identified as synonymous, present in more than four patients and considered as haplogroup markers were eliminated. "Confirmed" or "reported" variants in Mitomap or in literature and variants reported within the same region involved in a similar phenotype as the proband were analyzed. BWA, Burrows–Wheeler Aligner; GATK, Genome Analysis Toolkit; mtDNA, mitochondrial DNA; VCF, Variant Call Format

2.3.2 | Variants' calling, raw file generation, and filtering

Raw VCF files containing mtDNA variants were obtained for each individual with one GATK variant caller: HaplotypeCaller used with default parameters and which can detect simultaneously SNV and small indels. All raw VCF files were then merged in a global variant matrix file which grouped together with the data from all individuals. The first filter applied on the variant matrix consisted in keeping only rare variants, based on population frequency <1% (2012 Genomes Project Consortium et al., 2012) from the Mitomap database entries (Lott et al., 2013). As some of Leber's hereditary optic neuropathy (LHON) variants are detected with a 0.36% population frequency, we chose the arbitrary 1% value to be conservative. A supplementary filter was also applied for synonymous variants (Figures 1b and 2a).

The median number of individuals for the 2,429 raw variants was 2.0. Moreover, all well-known and "confirmed" mtDNA variants in mitochondrial disorders were checked. None was present in more than three individuals, so we chose to keep variants with less than four individuals.

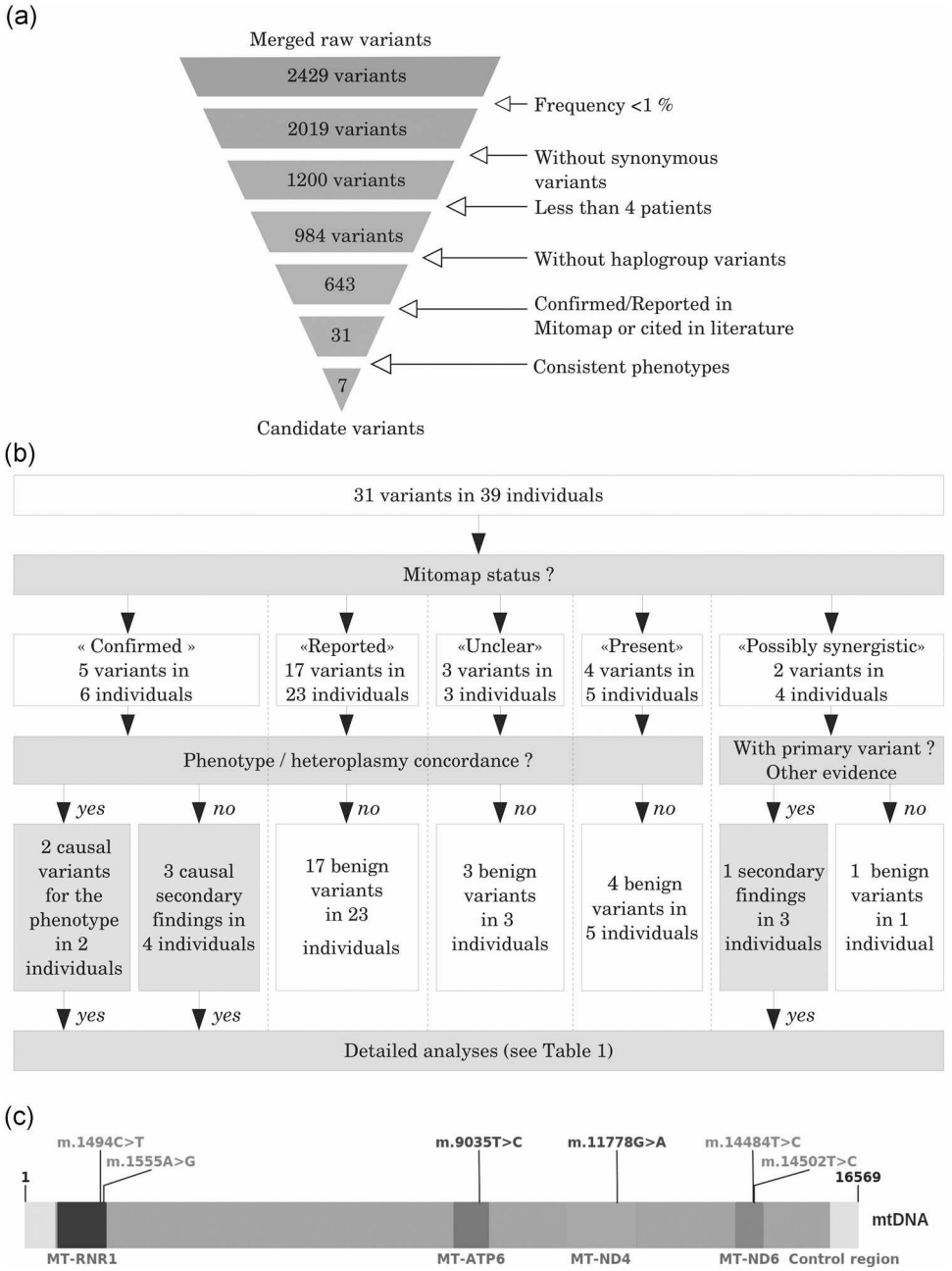


FIGURE 2 mtDNA variants of interest identification. (a) mtDNA variant filtering. Frequent variants and synonymous variants were filtered out at the beginning. Well represented variants in our database (present in more than four individuals) and variants defining haplogroups were removed. With these first four filters, 73.5% (n = 1,786) of variants were filtered out from the raw variant list, without considering the proband's phenotype. Looking for Mitomap status and citations in literature was not an automatic step as articles had to be analyzed and phenotypes compared. At the end of the process, 99.71% (n = 2,422) of variants were filtered out. (b) Mitomap variant filtering. Mitomap status was identified for each of the 31 variants (39 individuals). Each variant was then filtered by Mitomap frequency. Then, phenotype and/or heteroplasmy concordances were analyzed, except for "possibly synergistic" variants. These latter were filtered if no primary variant was present in involved individual mtDNA or if there was no other causal evidence. After this filtering step, only seven variants were detailed. (c) Identification of pathogenic mtDNA variants. The mtDNA is represented in linear form. Then the control region goes from 16024 to 576 (pink). Identified pathogenic variants are in blue and secondary findings in purple. All these variants are classified as "confirmed" or "reported—possibly synergistic" in Mitomap. *MT-RNR1*: 12S ribosomal RNA (MIM# 561000). *MT-ATP6* (MIM# 516060), *MT-ND4* (MIM# 516003), and *MT-ND6* (MIM# 516006) are protein-coding genes. The only regions in which pathogenic variants were identified are specified. mtDNA: mitochondrial DNA

2.3.3 | Haplogroup determination

In parallel, the mitochondrial haplogroup was determined from raw VCF files using HaploGrep 2.0 (Weissensteiner et al., 2016) online software. After individuals' haplogroups were determined, each variant involved in one or several haplogroups (Van Oven, 2015) was also annotated. When all individuals carrying the same variant had the same haplogroup or subhaplogroup, the variant was considered a polymorphism and filtered out. If the individuals did not have the same haplogroup, the considered variant was kept and further analyzed.

2.3.4 | Variant interpretation

The next analysis on the remaining variants consisted of checking the variant status in Mitomap (Lott et al., 2013; Figures 1b and 2b). We first retained the variants considered as "confirmed," "reported," or "unclear" in Mitomap. For the variants that were absent or present but not labeled, we explored other public databases (e.g., Clinvar [Landrum et al., 2018], OMIM [OMIM—Online Mendelian Inheritance in Man., 1996]) and scientific literature. Individuals' phenotypes and those described in Mitomap or in scientific articles were compared. When the status was "confirmed" or there was a probable concordance between the described phenotype and heteroplasmic rates, the variant was considered as a candidate variant and confirmed by PCR-RFLP or NGS. In cases of nonconcordance or absence in the databases, nearby regions were carefully examined for the presences of reported variants said to be involved in similar phenotypes.

Then, for tRNA variants with phenotype concordance, the Mitochondrial tRNA Informatics Predictors (MitoTIP) score was studied. The MitoTIP (Sonney et al., 2017) score provided by Mitomap is used to predict the probability that unknown or rare variants are pathogenic (score from 1%–99%). A tRNA variant is predicted as likely pathogenic or possibly pathogenic when its percentile score is more than 50%. On the contrary, with a score value less than 50%, the variant is predicted as possibly or likely benign.

To ensure that identified mtDNA possible causal variants were not due to NUMTs, nuclear equivalent regions were verified on mtDNA. The nuclear coordinates of NUMTs that were overlapping studied mtDNA sequences were extracted (Calabrese et al., 2012), and nuclear equivalent positions were manually checked.

2.3.5 | Depth analysis

Each position depth was determined by GATK DepthOfCoverage (Figure 4a). In parallel, NUMTs targeted during ES were obtained by intersecting exome capture kit target lists and the NUMT list (Calabrese et al., 2012) using BEDTools (Quinlan & Hall, 2010). Common nuclear regions were converted into mitochondrial coordinates and sorted by their chromosomal origins. Four Agilent SureSelect kits were used for ES and thus studied: Clinical Research Exome (CRE), CRE_v2, XT Human All Exon v4 and v5.

2.3.6 | Variant validation methods

A second molecular method (PCR, Sanger sequencing, or PCR-RFLP) was used to validate candidate variants (PCR and Sanger sequencing conditions are available in Supplementary Materials). The quantification of heteroplasmy was performed for each mtDNA variant using the fluorescent PCR-RFLP method. The region of interest was PCR amplified, the amplicons were digested, and the fragments produced were analyzed by capillary electrophoresis (Applied 3130XL). The results were analyzed with Peak Scanner Analysis Software (Thermo Fisher Scientific) and heteroplasmy was expressed as the percentage of mutant load.

2.3.7 | MToolBox tests

We also tested MToolBox on all positive cases and on our four positive controls to compare our method to an existing method.

The pipeline can be downloaded from http://gitlab.gad-bioinfo.org/gad-public/pipelinemito.

3 | RESULTS

All pathogenic mitochondrial SNV were identified in positive controls (Table S1), confirming the reliability of our method.

Among the 928 individuals, a total of 2429 mtDNA variants were obtained (Figure 2a). After frequency filtering (<1%), 2019 out of 2429 variants were kept. Almost half of them were extracted as nonsynonymous and 1200 out of 2019 underwent the rest of the process. Only 984/1200 variants were present in less than four individuals. After removing the variants defining haplogroups, 643/984 variants were analyzed. After we checked Mitomap and a completed thorough literature review, 31/643 variants in 39/928 individuals (4.2%) were considered candidate variants because they were found to have "confirmed," "reported," or "unclear" status (Table S4; Figure 2a).

The five "confirmed" Mitomap variants, NC_012920.1:m.1494C>T (*MT-RNR1*; MIM# 561000.0004; dbSNP Build 152: rs267606619), NC_012920.1:m.1555A>G (*MT-RNR1;* MIM# 561000.0001; dbSNP Build 152: rs267606617), NC_012920.1:m.9035T>C (YP_003024031.1:p.L170P; *MT-ATP6*; MIM# 516060), NC_012920.1:m.11778G>A (YP_003024035.1:p.R340H; *MT-ND4*; MIM# 516003.0001; dbSNP Build 152: rs199476112) and NC_012920.1:m.14484T>C (YP_003024037.1:p.M64V; *MT-ND6*; MIM# 516006.0001; dbSNP Build 152: rs199476104), were considered pathogenic in 6/39 individuals, contributing or responsible for individual features in 2/6 individuals, or as a secondary finding in 4/6 individuals. The almost homoplasmic m.9035T>C variant was confirmed by Sanger sequencing as heteroplasmic in blood and homoplasmic in urine in a 30-year-old female with learning disabilities, ataxia, and axonal neuropathy for

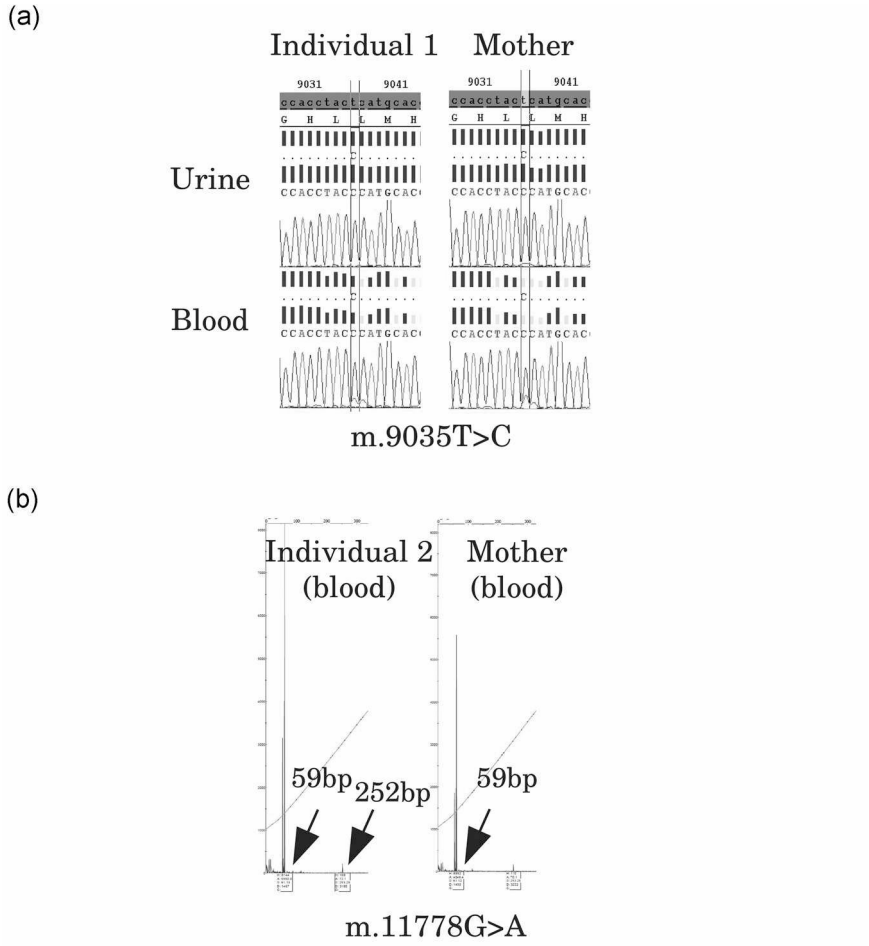


FIGURE 3 Molecular validation of individuals A and B candidate variants. (a) The m.9035T>C was confirmed by Sanger sequencing in blood and urine for both individual 1 and her mother. The mutant load was close to 100%. (b) The m.11778G>A was confirmed by PCR-RFLP in the blood and fibroblast cells of proband 2 and the blood of his mother. The heteroplasmic rate was almost 100% for proband 2, and the variant was homoplasmic in the blood of the mother. PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism

	Negative control (0%)	Positive control (22%)	Individual 2 (blood #1)	Individual 2 (blood #2)	Mother (blood)
59pb (variant)	0	11544	66518	66455	41533
252pb (WT)	51873	41370	1382	1342	0
Total	51873	52914	67900	67797	41533
Heteroplasmy %	0.0	21.82	97.96	98.02	100.00

10 years (Table 1 individual 1; Figure 3a) leading to clumsiness, fatigue, and balance problems. This variant was found in the same configuration in the individual's 59-year-old mother who had been diagnosed 6 years previously with isolated axonal neuropathy after undergoing family testing and had been experiencing pain and lower limb impairment (Figure 3a). The m.11778G>A variant was confirmed by PCR-RFLP in blood (heteroplasmic status) and fibroblasts (homoplasmic status) in a 30-year-old male, who presented a newly described mosaic neuroectodermal dysplasia, with dental, pigmentary, acral, and cerebral anomalies, eye malformation, and vision loss (Table 1 individual 2; Figure 3b). This mosaic condition had previously been explained by a causal postzygotic variant in a nuclear gene (Vabres et al., in 2019). However, the proband's severe ophthalmological impairment was atypical vis-à-vis the other individuals harboring pathogenic variants in the same gene; he was indeed the only individual in this cohort to carry the m.11778G>A variant (unpublished data) which was suspected of contributing to the ophthalmological severity. This variant was also found in the blood (homoplasmic) of his unaffected 59-year-old mother.

The three other "confirmed" in Mitomap variants were considered secondary findings identified in individuals without a matching phenotype (Table 1; Individuals 3–6). The m.1494C>T and m.1555A>G variants involved in aminoglycoside-induced deafness were found in a fetus with DA (cleft palate, oligodactyly, and lower limb hypoplasia; homoplasmic m.1494C>T variant), in one 19-year-old patient with ataxia (heteroplasmic m.1555A>G variant), and one 15-year-old patient with polymalformative syndrome (homoplasmic m.1555A>G variant). No clinical history of progressive deafness was

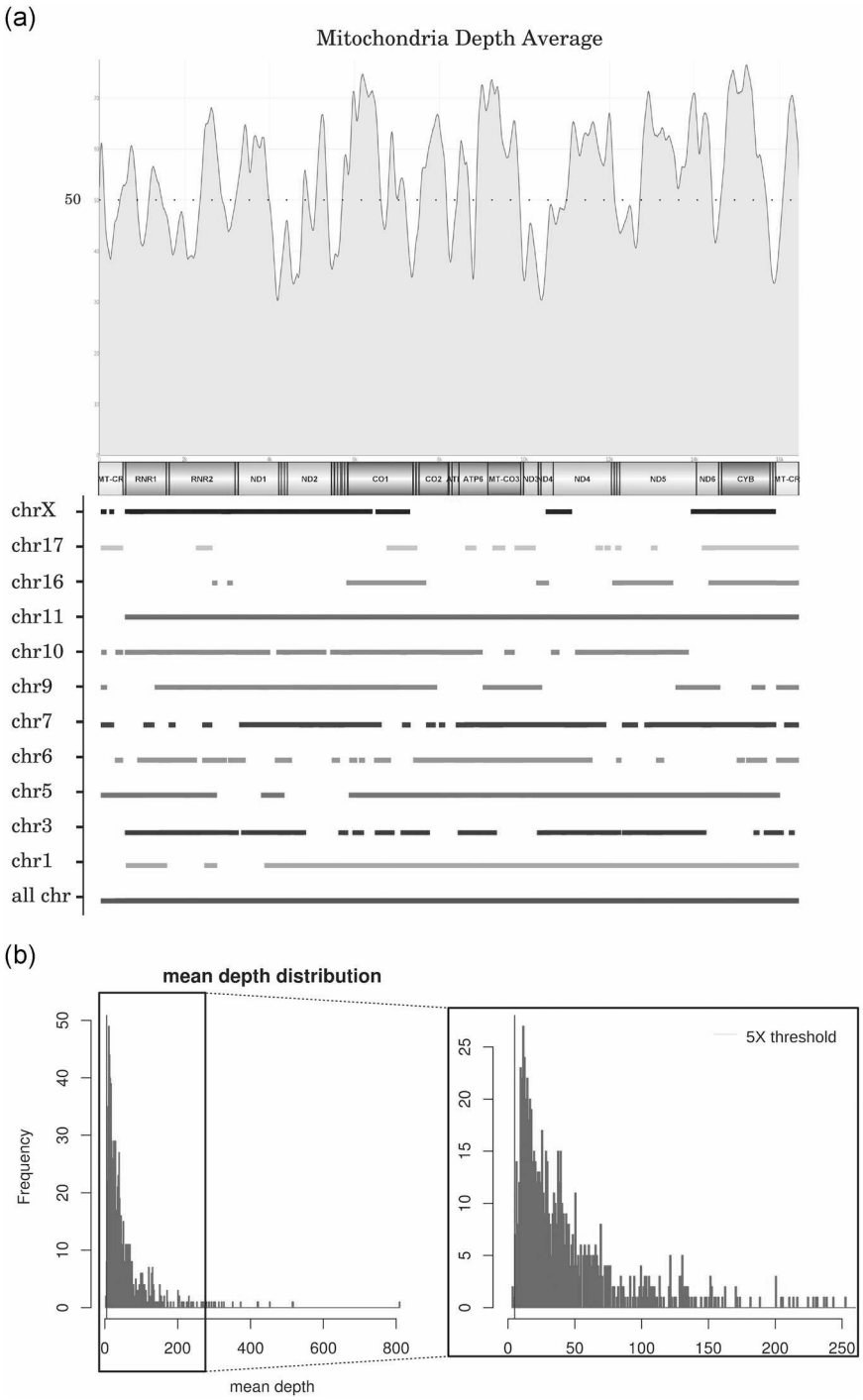


FIGURE 4 Off-target mitochondrial reads depth. (a) Smooth depth average is represented. All mitochondrial regions are covered with off-target reads. NUMTs are represented depending on their nuclear localization and their equivalent mitochondrial coordinates. (b) Mean depth distribution is presented. Threshold (×5) is represented by a blue vertical line. Less than 1% of the samples have a mean depth lower than ×5 (Griffin et al., 2014). NUMTs, nuclear mitochondrial DNA sequences

recorded in their family history. The homoplasmic m.14484T>C variant responsible for LHON (MIM# 535000) was identified in a 34-year-old individual suffering from muscular dystrophy but exhibiting no ophthalmological features. These variants were submitted to ClinVar, #SUB5620966 (https://www.ncbi.nlm.nih.gov/clinvar/).

Seventeen variants described as "reported" in Mitomap were identified in 23 individuals (Figure 2b). Sixteen were filtered out because the individuals' phenotype or heteroplasmic rates were not concordant with the literature. The homoplasmic NC_012920.1:m.5567T>C (*MT-TW*; MIM# 590095) variant was identified in two unrelated individuals (55 and 14 years old), both affected with cerebellar atrophy and ataxia and sharing the same haplogroup (K1a). The variant harbored a low MitoTIP score (32.70%), was predicted to be likely benign and identified in 41 GenBank sequences. The common mitochondrial background was also not in favor of its pathogenicity but could highlight a common evolving story. This variant was not kept for further analysis.

For the 11 remaining individuals, variants were filtered out because of nonconcordance between individual phenotypes or heteroplasmic rates and the literature. First, three variants found in three individuals were described as "unclear" in Mitomap or had conflicting reports. Two individuals with two of these variants were not considered because they were not consistent with the previously

TABLE 1 Individual characteristics. "Confirmed" and "reported—possibly synergistic" are variant statuses from Mitomap. Heteroplasmy rate was estimated by the ratio number of alternative reads/total number of reads covering the position. Haplogroup assignment was performed with HaploGrep 2.0 online software. The reference sequence used was GenBank NC_012920.1

Individual (sex)	Age (year)	Haplogroup determined (HaploGrep 2.0)	MtDNA variants	Mitomap status	Base depth	Described heteroplasmy rate	Individual heteroplasmy rate (alt/tot)	Variant associated diseases	Individual phenotype
1	30	U5a1i1	m.9035T>C	Confirmed	17	100%	100%	Ataxia syndromes (Pfeffer et al., 2012)	Progressive ataxia
(Female)			p.L170P			90–96%	17/17		
2	30	H6a1b3	m.11778 G>A	Confirmed	432	100%	100%	LHON (Wallace et al., 1988)	Ocular impairments + mosaic neuroectodermal dysplasia
(Male)			p.R340H			or less	432/432		
3	Fetus	H23	m.1494C>T	Confirmed	17	100%	100%	Aminoglycoside-induced non-syndromic deafness (Guan, 2004; H.Zhao et al., 2004)	Developmental anomalies
(Male)							17/17		
4	19	H1c	m.1555A>G	Confirmed	27	100% or less	66.67%	Aminoglycoside-induced non syndromic deafness (Casano et al., 1999; Fischel-Ghodsian, Prezant, Bu, & Öztas, 1993; Matsunaga et al., 2004)	Ataxia + nystagmus
(Male)							18/27		
5	15	H1be	m.1555A>G	Confirmed	15	100% or less	100%	Aminoglycoside-induced non-syndromic deafness (Casano et al., 1999; Fischel-Ghodsian et al., 1993; Matsunaga et al., 2004)	Polymalformative syndrome
(Male)							15/15		
6	34	X1′2′3	m.14484T>C	Confirmed	5	100% or less	100%	LHON (Wallace & Lott, 2017)	Muscular dystrophy
(Male)			p.M64V				5/5		
7	20	L2a1a2	m.14502T>C	Reported — possibly Synergistic	10	100%	100%	LHON (F.Zhao et al., 2009)	Rubinstein-Taybi syndrome
(Female)			p.I58V				10/10		
8	5	M10a1a1b1	m.14502T>C	Reported— possibly Synergistic	49	100%	97.95%	LHON (F.Zhao et al., 2009)	GLUT1 deficiency syndrome 1
(Female)			p.I58V				48/49		

(Continues)

TABLE 1 (Continued)

Individual (sex)	Age (year)	Haplogroup determined (HaploGrep 2.0)	MtDNA variants	Mitomap status	Base depth	Described heteroplasmy rate	Individual heteroplasmy rate (alt/tot)	Variant associated diseases	Individual phenotype
9	28	HV0a1	m.14502T>C	Reported—possibly Synergistic	3	100%	100%	LHON (F.Zhao et al., 2009)	Intellectual disabilities +developmental anomalies
(Female)			p.I58V				3/3		

Abbreviations: LHON, Leber's Hereditary Optic Neuropathy.
Note: The reference sequence used was GenBank NC_012920.1. These variants were submitted to ClinVar (SUB5620966): https://www.ncbi.nlm.nih.gov/clinvar/.
Haplogroup assignment was performed with HaploGrep 2.0 online software.

described phenotypes. The last individual, presenting sensorineural hearing loss due to the dilation of vestibules and aqueducts, had the homoplasmic NC_012920.1:m.8348A>G variant (*MT-TK*; MIM# 590060; dbSNP Build 152: rs1556423430). The phenotypes described for this variant diverged considerably, and its pathogenicity remained unclear: it was finally not considered pathogenic. Second, two variants, NC_012920.1:m.11696G>A (YP_003024035.1:p.V312I; *MT-ND4*; MIM# 516003; dbSNP Build 152: rs200873900) and NC_012920.1:m.14502T>C (YP_003024037.1:p.I58V; *MT-ND6*; MIM# 516006; dbSNP Build 152: rs201327354), found in four individuals were described as possibly synergistic when associated with primary pathogenic m.1555A>G (Deafness), m.11778G>A (LHON), and 14484T>C (LHON) variants. Synergistic variants are described as associated with primary pathogenic variants involved in known pathologies, thought to modulate the clinical phenotypes. There is some evidence that homoplasmic m.14502T>C could also be directly involved in LHON (F.Zhao et al., 2009) with lower penetrance than when it is associated with primary pathogenic variants. It was found in three individuals affected with Rubinstein–Taybi syndrome, GLUT1 deficiency syndrome and intellectual disability associated with DA. This homoplasmic m.14502T>C variant was considered to be a secondary finding and was confirmed with a second molecular method when DNA was still available. This variant was submitted to ClinVar, #SUB5620966 (https://www.ncbi.nlm.nih.gov/clinvar/). In the absence of primary pathogenic variants in their mtDNA and a nonconcordant phenotype, m.11696G>A variant was removed from the candidate variant list. The last four variants, found in five individuals, were already seen variants but not classified in Mitomap. These variants were not retained because the phenotypes were not concordant when compared to the literature data.

Most NUMTs appeared intergenic or intronic and therefore not covered by ES (unpublished data; Figure 4a). For possible causal variants, the study of nuclear NUMTs revealed that none of the variants was found in these regions, so the variants detected in our cohort were mtDNA variants as confirmed by molecular validation. In parallel, each position depth was studied: all the regions were covered by off-target reads but mean depth was variable. The NUMT analysis confirmed that the whole mtDNA could be captured with Agilent enrichment kits.

MToolBox was tested on 12 samples: our four positive controls, individuals 1-3, 5-7, 9, and the mother of the individual 2. This tool failed to identify one variant (m.14502T>C) in individual 9, confirmed by mtDNA NGS. Correlation coefficient value is thus 0.478 between MToolBox and specific mtDNA methods while it is 0.916 between our method and mtDNA specific methods.

4 | DISCUSSION

This study of indirect mtDNA sequencing on ES data in 928 individuals with DA and/or neurological disorders led to the identification of two different pathogenic variants (m.9035T>C and m.11778G>A) in 2/928 unrelated individuals (Table 1) responsible or

contributing to the phenotype (0.13% from DA cohort and 0.6% from neurological cohort) as well as secondary findings (m.1494C>T, m.1555A>G, m.14502T>C and m.14484T>C) in 7/928 unrelated individuals.

Certain bioinformatics or biological challenges were faced during the course of the study. The bioinformatics pipelines required substantial modifications to properly manage indirect mtDNA sequencing data (specific reference during alignment, specific databases, specific steps not existing in nuclear pipelines: mitochondrial chromosome extraction and haplogroup assignment). Indeed, although nuclear and mtDNA molecules were simultaneously extracted and sequenced (Samuels et al., 2013), nuclear and mitochondrial ES data could not be treated in the same way seeing as DNA references, databases and variant filters were different. The first improvement concerned the genome of reference. While most teams continued to align ES data on GRCh37/hg19, which is still mostly used in public databases, only GRCh38 contained the current mtDNA reference (Ye et al., 2014). It has been described as the best approach to decrease bias in heteroplasmic rate determination due to the overalignement of NUMTs. Since this study was based on off-target reads, it was essential to verify that detected mtDNA variants were not linked to NUMTs.

One of the major points of this study was the automatic haplogroup determination to improve variant interpretation and prioritization in a large unspecific cohort of patients not suspected of mitochondrial disorders. Haplogroup identification made it possible to verify the quality of mitochondrial genome reconstruction (Diroma et al., 2014). Haplogroup defining variants filtering for prioritization has already been described (Santorsola et al., 2016) in studies of tumor cells (F.M.Calabrese et al., 2016) or in a cohort of individuals suspected of mitochondrial diseases (Patowary et al., 2017).

In oncology, comparison of germline and somatic mtDNA variants highlighted specific mtDNA variants requiring prioritization filters such as population frequency or haplogroup assignment (F.M.Calabrese et al., 2016). In a cohort with suspected mitochondrial disorders, filtering out haplogroup defining variants made it possible to prioritize mtDNA variants, to confirm the causality of well-known variants for LHON (Santorsola et al., 2016) during performance evaluation of the prioritization criteria, or to highlight variants of interest in autism (Patowary et al., 2017), a disease, which may be linked to mitochondrial dysfunction. In our large cohort, which was not suspected of mitochondrial disorders, haplogroup determination removed common polymorphisms not filtered upstream.

Variant interpretation also presented "biological" challenges and limits. The first group of five "confirmed" variants in Mitomap was straightforward because the variants were well described and their pathogenicity demonstrated. When phenotype and mutant load were concordant with the literature, the variants were considered as causal after molecular validation (Individuals 1 and 2). When phenotypes were nonconcordant with those previously described in the literature (Individuals 3–9), variants were then considered as secondary findings, given that incomplete penetrance was described for mtDNA variants. Individual 1 and her mother carried the homoplasmic m.9035T>C variant, which was responsible for ataxia and a milder maternal phenotype (Figure 3a). This variant has previously been reported in ataxic syndromes (Pfeffer et al., 2012). Individual 2 and his unaffected mother carried the homoplasmic m.11778G>A variant (Figure 3b), associated with low penetrance LHON (Wallace et al., 1988). This molecular double hit allowed us to specify the phenotypical spectrum of the neuroectodermal disorder, providing important data for genetic counseling.

Other variants in Mitomap were classified as "reported," "synergistic," and "unclear" or were unclassified variants. Their pathogenicity was not clearly established and excluded because of the absence of primary variants, low pathogenic prediction score, or phenotype nonconcordance.

One of the well-known limits in the exploration of mtDNA is the mutational rate since the heteroplasmic state depends heavily on the choice of tissues. As mtDNA data derived from off-target ES sequences, mtDNA regions were not homogeneously covered (Figure 4a). The mean depth was lower in mtDNA (×50 mean coverage) compared with specifically targeted nuclear genes (×100 mean coverage). The differences may be due to the capture kits chosen for ES, since each kit presented with a distinct design and different targeted sequences. Thus, off-target region sequencing differed, resulting in nonuniform mitochondrial coverage. Some variants or heteroplasmic rates could not be detected because depth and coverage obtained by this method are not optimal for mitochondria study and significantly below mtDNA NGS study (~×500; Figure 4b). Specific mtDNA analysis remains therefore indicated for negative results after our method. Nevertheless, as shown with positive controls analysis, our method can both determine homoplasmic/heteroplasmic status and heteroplasmic rate even with low depth. Heteroplasmic rates were comparable between the initial method and our determination (correlation coefficient 0.943). Moreover, a good correlation between mtDNA NGS data and our sequencing data can be observed, leading to an increase of diagnostic yield as new cases have been solved. As GS can use PCR free technology, the adaptation of our method to a GS pipeline would obtain better coverage and depth. Indeed, mtDNA molecules are 10 to 100 times more present in cells than nDNA molecules (Dinwiddie et al., 2013) improving the identification of variants and heteroplasmic rates. A lower depth made the interpretation of variants difficult, considering the presence of a variant calling threshold.

The determination of the mutant load is also impacted depending on the number of aligned reads (Griffin et al., 2014). For this study, ES data were extracted from only one type of tissue, mostly blood, in which a candidate variant could be absent, at a low level or even undetectable. As the coverage remains heterogeneous, NGS analysis of the whole mtDNA still remains indicated in individuals highly suspected of being affected by a mitochondrial disorder.

Incomplete penetrance is another challenge for variant interpretation. Indeed, even if family segregation can help, it can remain difficult to diagnose a mitochondrial disorder because of incomplete penetrance or unusual phenotypical variability.

We did not expect the rate of secondary findings to be higher than the rate of a positive diagnosis. These molecular findings are responsible for well-defined phenotypes with incomplete penetrance, estimated at 0.53% patients with LHON in this cohort, more than the expected rate in the general population of 0.29% (1/350), and 0.3% of patients with aminoglycoside-induced deafness, similar to the expected rate in the general population (0.28%). It is worth discussing whether these results should be returned to patients considering their potential use for health prevention counseling.

To date, only one study has reported the development of analytical tools for evaluating mtDNA in whole-exome data for the diagnosis of rare diseases (Patowary et al., 2017). Patowary et al. (2017) studied ten multiplex families with autism to provide further support for the role of mitochondria in autism spectrum disorders. They confirmed that whole-exome sequencing may be combined with mtDNA analysis. They highlighted the challenges of analyzing mtDNA variants with ES, including phenotype heterogeneity, age of onset, incomplete penetrance, or the heteroplasmic rate. They chose MToolBox to detect mitochondrial variants from ES data. This tool used a different method to analyze mtDNA sequence. ES data are first aligned on mitochondrial reference before hg19 reference. Reads uniquely mapped on mtDNA reference are kept. They are then used to generate a VCF file, to determine heteroplasmy rate, and to reconstruct a complete mitochondrial genome. Based on this latter, an individual haplogroup is assigned by macro-haplogroup-specific consensus sequence alignment. Our method does not need reconstructed mitochondrial genome to assign haplogroup thanks to HaploGrep2 tool. In addition, allele frequencies were extracted from 1,000 genomes samples rather than from Mitomap, a specific mtDNA database also containing more entries.

The correlation coefficient between our method and specific mtDNA methods is higher than the correlation coefficient between MtoolBox and specific mtDNA methods. Our pipeline can thus be considered as more sensitive.

In conclusion, we developed a bioinformatics pipeline to prospectively identify mtDNA variants by indirect mtDNA sequencing from ES data parallel to nuclear exome analysis. After testing the approach on a series of patients carrying pathogenic mtDNA variants, we confirmed the interest of implementing this approach systematically in routine ES bioinformatics pipelines in large cohorts. This technique provides a significant opportunity to reduce the diagnostic odyssey (Thevenon et al., 2016) for certain patients. However, the question of how to manage secondary findings from mtDNA remains complex and warrants further discussion.

ACKNOWLEDGMENT

We thank the probands and their families for their participation; and the Center De Calcul (CCuB) at the University of Burgundy for providing technical support and management of the informatics core facility. This work was supported by grants from the Regional Council of Burgundy (to C.T.-R.), the FEDER 2017, PARI 2017, and CIFRE (ANRT) between Laboratoire Cerba and Regional Council of Burgundy for the doctoral work at Laboratoire Cerba and GAD.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ORCID

Philippine Garret http://orcid.org/0000-0003-2551-3606
Céline Bris http://orcid.org/0000-0003-2425-4446
Vincent Procaccio http://orcid.org/0000-0002-1537-4684
Pierre Vabres http://orcid.org/0000-0001-8693-3183
François Feillet http://orcid.org/0000-0002-6814-0806
Ange-Line Bruel http://orcid.org/0000-0002-0526-465X
Virginie Quéré http://orcid.org/0000-0001-8802-6448
Christophe Philippe http://orcid.org/0000-0001-7098-6520
Arthur Sorlin http://orcid.org/0000-0001-8008-9145
Antonio Vitobello http://orcid.org/0000-0003-3717-8374
Laurence Faivre http://orcid.org/0000-0001-9770-444X
Christel Thauvin-Robinet http://orcid.org/0000-0002-4155-139X

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, *23*(2), 147. https://doi.org/10.1038/13779

Bai, R. K., & Wong, L. J. C. (2005). Simultaneous detection and quantification of mitochondrial DNA deletion(s), depletion, and over-replication in patients with mitochondrial disease. *The Journal of Molecular Diagnostics*, *7*(5), 613–622. https://doi.org/10.1016/S1525-1578(10)60595-8

Bannwarth, S., Procaccio, V., Lebre, A. S., Jardel, C., Chaussenot, A., Hoarau, C., … Paquis-Flucklinger, V. (2013). Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders. *Journal of Medical Genetics*, *50*, 704–714. https://doi.org/10.1136/jmedgenet-2013-101604. jmedgenet-2013-101604.

Broad Institute. (n.d). Picard Tools. Retrieved May 2016, from http://broadinstitute.github.io/picard/

Calabrese, C., Simone, D., Diroma, M. A., Santorsola, M., Guttà, C., Gasparre, G., … Attimonelli, M. (2014). MToolBox: A highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, *30*(21), 3115–3117. https://doi.org/10.1093/bioinformatics/btu483.

Calabrese, F. M., Clima, R., Pignataro, P., Lasorsa, V. A., Hogarty, M. D., Castellano, A., … Capasso, M. (2016). A comprehensive characterization of rare mitochondrial DNA variants in neuroblastoma. *Oncotarget*, *7*(31), 49246–49258. https://doi.org/10.18632/oncotarget.10271

Calabrese, F. M., Simone, D., & Attimonelli, M. (2012). Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics*, *13*(Suppl 4), S15. https://doi.org/10.1186/1471-2105-13-S4-S15

Casano, R. A. M. S., Johnson, D. F., Bykhovskaya, Y., Torricelli, F., Bigozzi, M., & Fischel-Ghodsian, N. (1999). Inherited susceptibility to aminoglycoside

ototoxicity: Genetic heterogeneity and clinical implications. *American Journal of Otolaryngology*, *20*(3), 151–156. https://doi.org/10.1016/S0196-0709(99)90062-5

Chinnery, P. F. (2002). Inheritance of mitochondrial disorders. *Mitochondrion*, *2*(1–2), 149–155. https://doi.org/10.1016/S1567-7249(02)00046-6

Coulter-Mackie, M. B., Applegarth, D. A., Toone, J. R., & Gagnier, L. (1998). A protocol for detection of mitochondrial DNA deletions: Characterization of a novel deletion. *Clinical Biochemistry*, *31*(8), 627–632. https://doi.org/10.1016/S0009-9120(98)00074-5

Dinwiddie, D. L., Smith, L. D., Miller, N. A., Atherton, A. M., Farrow, E. G., Strenk, M. E., … Kingsmore, S. F. (2013). Diagnosis of mitochondrial disorders by concomitant next-generation sequencing of the exome and mitochondrial genome. *Genomics*, *102*, 148–156. https://doi.org/10.1016/j.ygeno.2013.04.013

Diroma, M. A., Calabrese, C., Simone, D., Santorsola, M., Calabrese, F. M., Gasparre, G., & Attimonelli, M. (2014). Extraction and annotation of human mitochondrial genomes from 1000 genomes whole exome sequencing data. *BMC Genomics*, *15*(Suppl 3), S2. https://doi.org/10.1186/1471-2164-15-S3-S2

Falk, M. J., Pierce, E. A., Consugar, M., Xie, M. H., Guadalupe, M., Hardy, O., … Gai, X. (2012). Mitochondrial disease genetic diagnostics: Optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome. *Discovery Medicine*, *14*(79), 389–399.

Fischel-Ghodsian, N., Prezant, T. R., Bu, X., & Öztas, S. (1993). Mitochondrial ribosomal RNA gene mutation in a patient with sporadic aminoglycoside ototoxicity. *American Journal of Otolaryngology*, *14*(6), 399–403. https://doi.org/10.1016/0196-0709(93)90113-L

Gorman, G. S., Chinnery, P. F., DiMauro, S., Hirano, M., Koga, Y., McFarland, R., … Turnbull, D. M. (2016). Mitochondrial diseases. *Nature Reviews Disease Primers*, *2*, 16080. https://doi.org/10.1038/nrdp.2016.80

Griffin, H. R., Pyle, A., Blakely, E. L., Alston, C. L., Duff, J., Hudson, G., … Chinnery, P. F. (2014). Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. *Genetics In Medicine*, *16*(12), 962–971. https://doi.org/10.1038/gim.2014.66

Guan, M. -X. (2004). Molecular pathogenetic mechanism of maternally inherited deafness. *Annals of the New York Academy of Sciences*, *1011*(1), 259–271. https://doi.org/10.1196/annals.1293.025

Guo, Y., Li, J., Li, C.-I., Shyr, Y., & Samuels, D. C. (2013). MitoSeek: Extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics*, *29*(9), 1210–1211. https://doi.org/10.1093/bioinformatics/btt118

He, L., Chinnery, P. F., Durham, S. E., Blakely, E. L., Wardell, T. M., Borthwick, G. M., … Turnbull, D. M. (2002). Detection and quantification of mitochondrial DNA deletions in individual cells by real-time PCR. *Nucleic Acids Research*, *30*(14), e68.

He, Y., Wu, J., Dressman, D. C., Iacobuzio-Donahue, C., Markowitz, S. D., Velculescu, V. E., … Papadopoulos, N. (2010). Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, *464*(7288), 610–614. https://doi.org/10.1038/nature08802

Holt, I. J., Harding, A. E., Petty, R. K., & Morgan-Hughes, J. A. (1990). A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *American Journal of Human Genetics*, *46*, 428–433.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(Database issue), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., … Wallace, D. C. (2013). mtDNAVariation and analysis using mitomap and mitomaster. *Current Protocols in Bioinformatics*, *44*(1), 1.23.1–1.23.26. https://doi.org/10.1002/0471250953.bi0123s44

Matsunaga, T., Kumanomido, H., Shiroma, M., Ohtsuka, A., Asamura, K., & Usami, S. (2004). Deafness due to A1555G mitochondrial mutation without use of aminoglycoside. *The Laryngoscope*, *114*(6), 1085–1091. https://doi.org/10.1097/00005537-200406000-00024

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A map reduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Online Mendelian Inheritance in Man (OMIM) (1996). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Retrieved from https://www.omim.org/

Patowary, A., Nesbitt, R., Archer, M., Bernier, R., & Brkanac, Z. (2017). Next generation sequencing mitochondrial DNA analysis in autism spectrum disorder. *Autism Research*, *10*(8), 1338–1343. https://doi.org/10.1002/aur.1792

Pfeffer, G., Blakely, E. L., Alston, C. L., Hassani, A., Boggild, M., Horvath, R., … Chinnery, P. F. (2012). Adult-onset spinocerebellar ataxia syndromes due to MTATP6 mutations. *Journal of Neurology, Neurosurgery & Psychiatry*, *83*(9), 883–886. https://doi.org/10.1136/jnnp-2012-302568

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. *Trends in Genetics*, *29*(10), 593–599. https://doi.org/10.1016/j.tig.2013.07.006

Santorsola, M., Calabrese, C., Girolimetti, G., Diroma, M. A., Gasparre, G., & Attimonelli, M. (2016). A multi-parametric workflow for the prioritization of mitochondrial DNA variants of clinical interest. *Human Genetics*, *135*, 121–136. https://doi.org/10.1007/s00439-015-1615-9

Sherry, S. T., Ward, M. -H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Sonney, S., Leipzig, J., Lott, M. T., Zhang, S., Procaccio, V., Wallace, D. C., & Sondheimer, N. (2017). Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. *PLOS Computational Biology*, *13*(12), e1005867.

Thevenon, J., Duffourd, Y., Masurel-Paulet, A., Lefebvre, M., Feillet, F., El chehadeh-Djebbar, S., … Rivière, J. B. (2016). Diagnostic odyssey in severe neurodevelopmental disorders: Toward clinical whole-exome sequencing as a first-line diagnostic test. *Clinical Genetics*, *89*(6), 700–707. https://doi.org/10.1111/cge.12732

Vabres, P., Sorlin, A., Kholmanskikh, S. S., Demeer, B., St-Onge, J., Duffourd, Y., … Rivière, J.-B. (2019). Postzygotic inactivating mutations of RHOA cause a mosaic neuroectodermal syndrome. *Nature Genetics*.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). Current protocols in bioinformatics: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *11*(1110), 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Van Oven, M. (2015). PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International*, *5*, e392–e394. https://doi.org/10.1016/j.fsigss.2015.09.155

Vasta, V., Ng, S. B., Turner, E. H., Shendure, J., & Hahn, S. H. (2009). Next generation sequence analysis for mitochondrial disorders. *Genome Medicine*, *1*(10), 100. https://doi.org/10.1186/gm100

Vellarikkal, S. K., Dhiman, H., Joshi, K., Hasija, Y., Sivasubbu, S., & Scaria, V. (2015). mit-o-matic: A comprehensive computational pipeline for clinical evaluation of mitochondrial variations from next-generation sequencing datasets. *Human Mutation*, *36*(4), 419–424. https://doi.org/10.1002/humu.22767

Wallace, D., Singh, G., Lott, M., Hodge, J., Schurr, T., Lezza, A., … Nikoskelainen, E. (1988). Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science*, *242*(4884), 1427–1430.

Wallace, D. C., Fan, W., & Procaccio, V. (2010). Mitochondrial energetics and therapeutics. *Annual Review of Pathology*, *5*, 297–348. https://doi.org/10.1146/annurev.pathol.4.110807.092314

Wallace, D. C., & Lott, M. T. (2017). Leber hereditary optic neuropathy: Exemplar of an mtDNA disease. In Singh, H., & Sheu, S. -S. (Eds.), *Pharmacology of Mitochondria* (240, pp. 339–376). Cham: Springer International Publishing. https://doi.org/10.1007/164_2017_2

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., … Schönherr, S. (2016). HaploGrep 2: Mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research*, *44*(W1), W58–W63. https://doi.org/10.1093/nar/gkw233

Ye, F., Samuels, D. C., Clark, T., & Guo, Y. (2014). High-throughput sequencing in mitochondrial DNA research. *Mitochondrion*, *17*, 157–163. https://doi.org/10.1016/j.mito.2014.05.004

Zhang, P., Samuels, D. C., Lehmann, B., Stricker, T., Pietenpol, J., Shyr, Y., & Guo, Y. (2016). Mitochondria sequence mapping strategies and practicability of mitochondria variant detection from exome and RNA sequencing data. *Briefings in Bioinformatics*, *17*(2), 224–232. https://doi.org/10.1093/bib/bbv057

Zhao, F., Guan, M., Zhou, X., Yuan, M., Liang, M., Liu, Q., … Guan, M. X. (2009). Leber's hereditary optic neuropathy is associated with mitochondrial ND6 T14502C mutation. *Biochemical and Biophysical Research Communications*, *389*(3), 466–472. https://doi.org/10.1016/j.bbrc.2009.08.168

Zhao, H., Li, R., Wang, Q., Yan, Q., Deng, J. -H., Han, D., … Guan, M. X. (2004). Maternally inherited aminoglycoside-induced and nonsyndromic deafness is associated with the novel C1494T mutation in the mitochondrial 12S rRNA gene in a large chinese family. *The American Journal of Human Genetics*, *74*(1), 139–152. https://doi.org/10.1086/381133

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Garret P, Bris C, Procaccio V, et al. Deciphering exome sequencing data: Bringing mitochondrial DNA variants to light. *Human Mutation*. 2019;40:2430–2443. https://doi.org/10.1002/humu.23885

TROISIÈME PARTIE : IDENTIFICATION DES ÉLÉMENTS MOBILES À PARTIR DE DONNÉES DE SÉQUENÇAGE D'EXOME

I- INTRODUCTION

Les éléments mobiles, absents de la référence génomique en cas d'insertion récente, ne sont pas aisément détectables avec une méthode « classique » d'analyse d'exome ou de génome car ils présentent des séquences répétées ou génèrent des lectures multimappées. Les logiciels de détection de ces EM utilisent les paires de lectures discordantes (DP) : une des lectures est alignée sur le génome de référence tandis que l'autre s'aligne sur un élément mobile (**Fig. 39a**). Ils permettent également d'identifier le point de cassure.

Les lectures sont alignées à la fois sur le génome de référence et sur une base de données d'éléments mobiles. Les paires de lectures s'alignant uniquement sur un EM ne sont pas retenues. Les programmes commencent par détecter les DP. En fonction des logiciels, peuvent également être détectées les paires de lectures contenant une lecture « splittée » (SR) (**Fig. 39b**). Il s'agit de paires constituées d'une lecture alignée de façon unique sur la référence ainsi que d'une lecture qui s'aligne par une extrémité sur la référence génomique et par l'autre extrémité sur la base de données des EM. Certains programmes y ajoutent aussi les paires avec une lecture alignée sur le génome de référence et une lecture non alignée. Les régions du génome qui présentent des alignements DP et/ou SR sont suspectées de contenir un EM inséré. Elles seront filtrées par les différents programmes en fonction du nombre de DP et/ou SR permettant leur mise en évidence (valeur seuil définie dans les paramètres par l'utilisateur).

Les DP permettent de délimiter la région d'insertion de l'EM. Puis les lectures qui s'alignent sur le génome de référence sont regroupées en deux clusters distincts, en 5' ou en 3' en fonction de leur localisation par rapport à l'EM suspecté. Les bornes internes de ces clusters représentent alors les frontières de la région d'insertion de l'EM. Les SR permettent ensuite de déterminer la position exacte du point de cassure sur le génome. En effet, l'alignement des SR sur le génome de référence puis sur la base de données d'EM permet de mettre en évidence la jonction référence-EM sur la lecture et de déterminer avec précision le point de cassure.

Le recoupement de ces données permet de déterminer la position du point de cassure sur le génome ainsi que la nature de l'EM inséré (L1, Alu ou SVA).

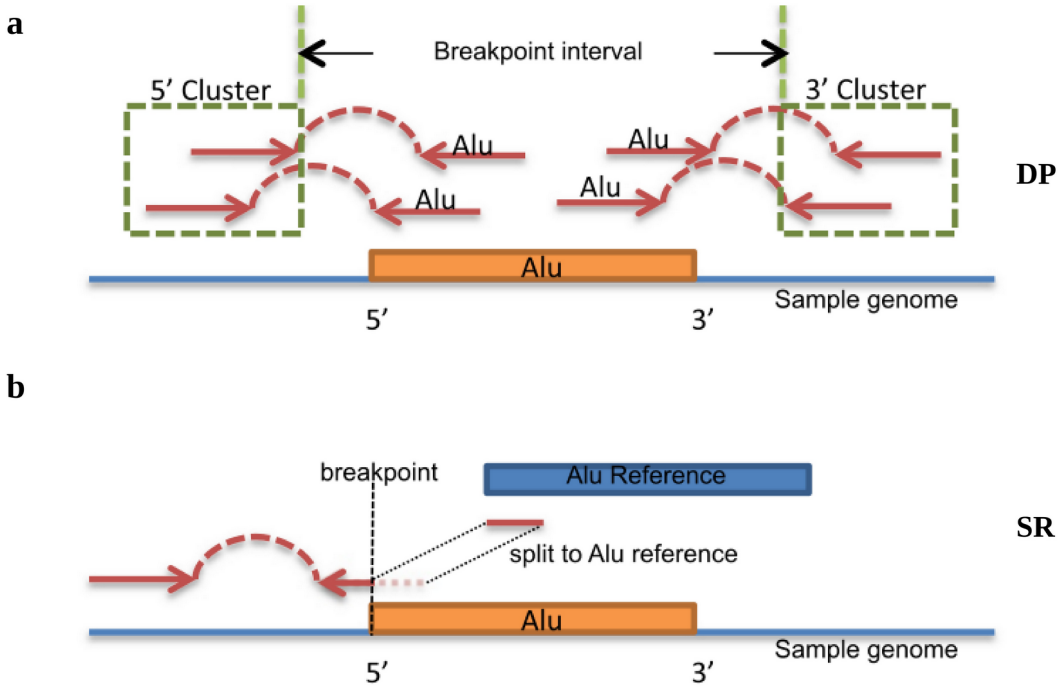


Figure 39 : Principes de détection des EM à partir des données de séquençage à haut débit *Identification de la région d'insertion de l'élément mobile (orange) à partir des paires de lectures discordantes DP ; et du point de cassure à partir des lectures « splittées » SR (d'après (Wu et al., 2014)).*

II- MATÉRIEL ET MÉTHODES

II.1- Contrôles positifs et cohorte de patients

Les 3 pipelines d'identification des EM ont tout d'abord été testés sur des données de séquençage de génome de 29 individus contrôles positifs du projet 1000 Genomes choisi aléatoirement. En effet l'identification *in silico* d'EM non référencés a déjà été entreprise sur ces données et l'existence de ces EM a été biologiquement confirmée (Stewart et al., 2011 ; 1000 Genomes Project, 2015). Il s'agissait de tester la faisabilité de cette étude d'un point de vue informatique et de vérifier la détection des EM par 3 programmes. Les effectifs de chaque

EM (Alu, L1 et SVA) pour chaque individu ont été comparés avec les valeurs obtenues par le consortium 1000 Genomes (1000 Genomes Project, 2015) qui a utilisé l’outil MELT.

L’identification des EM a été réalisée sur une cohorte de 3322 individus séquencés en exome dont 2500 cas index. Environ 80 % des patients étaient atteints d’anomalies du développement avec ou sans atteinte neurologique et ~20 % présentaient une maladie neurologique rare. Environ 25 % des exomes avaient un résultat positif en analyse diagnostique. La proportion homme-femme était de 50 %.
Un séquençage à haut débit d'exome en solo ou trio à partir des échantillons ADN a été réalisé chez chaque patient. La capture a été effectuée avec un kit de chez Agilent : v3, v4, v5, v6, v7, Clinical Research Exome, Clinical Research Exome v2 ou plus récemment avec le kit TWIG. Ce dernier correspond au kit Human Core Exome de chez Twist Bioscience complété avec des sondes supplémentaires. Le séquençage a été réalisé soit sur un HiSeq 2000, HiSeq 4000, NextSeq 550 ou NovaSeq 6000 (Integragen) avec une lecture en paired-end et des lectures de 100 pb. Les fichiers d’alignement (BAM) ont été obtenus selon le protocole décrit en partie II.1.2.
La cohorte des 433 individus ayant été séquencés en génome comportaient 165 cas index avec une proportion homme-femme de 50 %. Pour environ 18 % des patients le diagnostic moléculaire avait été posé.

II.2- Outils bioinformatiques testés au cours du projet

Afin de détecter les éléments mobiles actifs (Alu, L1 et SVA) à partir des données de génome ou d’exome d’une cohorte de 3322 individus, 3 logiciels de détection ont été utilisés. Les résultats de l’outil MELT ont été analysés et les candidats retenus ont été vérifiés dans les résultats des outils Mobster et MELT.

Le programme Tangram a été publié en 2014 (Wu et al., 2014). Validé sur les données de 1000 génomes, il fonctionne selon le principe de détection des DP et SR après un alignement

sur le génome de référence et sur 23 références d'éléments mobiles (17 L1, 4 Alu, 1 SVA et 1 HERV-K) provenant de la base de données RepBase (Bao et al., 2015) par le programme d'alignement MOSAIK (Lee et al., 2014b). Il conserve les candidats ayant au minimum 2 DP en 5' et en 3' ou 2 SR. En l'absence de SR, le candidat n'est pas retenu s'il se trouve à proximité d'un locus présent au sein de la base de données RepeatMasker (Smit et al., 2008). Partant de données sous la forme de fichiers BAM il génère un VCF donnant la position de l'EM, son type et le génotype du patient étudié.

Le programme Mobster a été publié en 2014 (Thung et al., 2014). Ce logiciel détecte de nouveaux éléments mobiles. Il fait appel à une base de données, appelée mobilome, composée de 54 séquences consensus d'éléments mobiles provenant de la base de données RepBase 17.3. Le programme fonctionne à partir de DP et de lectures clippées. Mobster s'exécute sur des fichiers BAM d'alignement produits par BWA ou MOSAIK sur des données de séquençage d'exome ou de génome puis génère un VCF.

Le programme MELT (Mobile Element Locator Tool) a été publié en 2017 (Gardner et al., 2017). Il détecte les éléments Alu, L1 et SVA par identification des DP puis des lectures chevauchantes au sein de données de génome (BAM) alignées sur le génome de référence par BWA et sur les séquences de référence des éléments mobiles alignées par Bowtie 2. MELT récupère également des informations sur l'EM identifié : chromosome, orientation, séquence du TSD, type, taille, génotype du patient, etc. Il indique également dans le VCF final les gènes concernés par l'insertion, l'identifiant RefSeq du transcrit, le segment du gène (exon, intron, promoteur, etc.). Enfin il effectue des calculs de scores afin d'estimer la qualité de la prédiction. Les prédictions sont filtrées par la profondeur de séquençage, la qualité d'alignement des lectures, leur localisation et leur proximité avec les séquences de référence des EM.

Afin de faciliter l'analyse des éléments mobiles identifiés, une annotation du fichier VCF contenant les positions génomiques de ces EM est réalisée à l'aide du logiciel AnnotSV (Geoffroy et al., 2018). Ce dernier a été conçu afin d'annoter et de prioriser les SV du génome humain. Ces annotations concernent notamment les informations sur les gènes RefSeq ou les régions génomiques impliquées, présentes au sein de nombreuses bases de données :

- nom du gène, identifiant RefSeq du transcrit, score pLI, score MISZ et score d'haploinsuffisance du gène, numéro OMIM, phénotype et hérédité en cas de gènes

morbides.

- annotations par les bases de données de variations structurales Deciphering Developmental Disorders (DDD), Database of Genomic Variants (DGV) (hérédité, pathologie, phénotype, fréquence, etc.), dbVar, 1000 Genomes.
- identification de régions promotrices touchées, des TAD, des séquences répétées (type et coordonnées génomiques).
- calcul du contenu en GC en 5' et en 3' du point de cassure.

AnnotSV nécessite au minimum un fichier BED ou VCF contenant les coordonnées génomiques des SV/EM à annoter. Il produit un fichier tabulé contenant des annotations de deux types : « full » et « split ». La première annote tous les éléments mobiles y compris ceux ne touchant pas de gènes. La seconde annote en plus des éléments mobiles les gènes concernés.

Pour cette étude, l'identification des éléments mobiles s'est basée sur les résultats obtenus avec le programme MELT. Les candidats retenus ont ensuite été recherchés dans les résultats des programmes Tangram et Mobster.

II.4.1- Pipeline MELT

Les 3322 exomes d'individus ont été divisés en 8 groupes d'analyse en respectant les trios lorsque les exomes des parents ou d'autres membres de la famille étaient disponibles. Le pipeline d'identification des EM avec l'outil MELT a été implémenté pour analyser des données d'exome en « batch » : stratégie d'analyse de plusieurs exomes en simultané. Ces données ont été alignées avec BWA (**Fig. 40**). Avant l'analyse, la profondeur de l'exome de chaque individu a dû être calculée et le fichier BAM préprocessé selon la méthode décrite plus haut (voir la section « Traitement et alignement des données brutes de séquençage d'exome »). La profondeur moyenne d'un exome a donc été calculée à l'aide d'un script python se basant sur les données obtenues par Samtools. Cette valeur a été utilisée par MELT pour l'étape de preprocessing. L'étape suivante a consisté en la découverte d'EM chez chaque individu (étape « IndivAnalysis »). Puis le programme a réuni toutes les données du groupe afin d'obtenir plus

de précision concernant notamment les points de cassures, la séquence TSD, le sens, la taille, la famille des éléments mobiles détectés (« GroupAnalysis »). MELT a ensuite déterminé le génotype de chaque individu pour les éléments mobiles détectés (« Genotype »). Enfin le programme a filtré les résultats et les a réunit dans un fichier VCF commun à tous les individus du groupe (« MakeVCF »). Trois fichiers VCF ont ainsi été obtenus : un pour chaque type d'élément mobile. Chaque famille de patient analysée est décrite dans un fichier regroupant les identifiants de ses membres. De plus, pour chaque individu du groupe il existe un fichier contenant son statut dans la famille (cas index, père, mère, etc.). Après le traitement par MELT, le pipeline a ainsi pu déterminer pour chaque cas index s'il s'agissait d'une analyse en solo ou en trio (cas index + famille). Il a ensuite extrait les données du ou des individus de la famille en générant un fichier VCF par cas index et par élément mobile. Ces fichiers contiennent le génotype du cas index et des parents le cas échéant. Pour chaque cas index, les 3 fichiers VCF ont ensuite été concaténés. Puis le fichier VCF unique a été annoté par le programme AnnotSV. Un rapport final a alors été généré par un script python pour chaque cas index du groupe. Il s'agit d'un fichier tabulé contenant les informations présentes au sein du VCF, ce format tabulé permet une meilleure lisibilité pour l'interpréteur.

A l'issue de l'analyse bioinformatique des 8 groupes, les éléments mobiles ont été filtrés pour ne retenir que ceux présents au sein d'un gène OMIM morbide, d'une région non-intronique (5'- et 3'-UTR, exoniques, régions promotrices et terminateurs), présents moins de 5 fois par groupe et dans la cohorte des 3322 individus. Un script python a ensuite regroupé tous les éléments mobiles ayant passé ces filtres. L'identifiant de l'échantillon et la profondeur au sein du BAM d'origine au niveau du point de cassure ont été ajoutés.

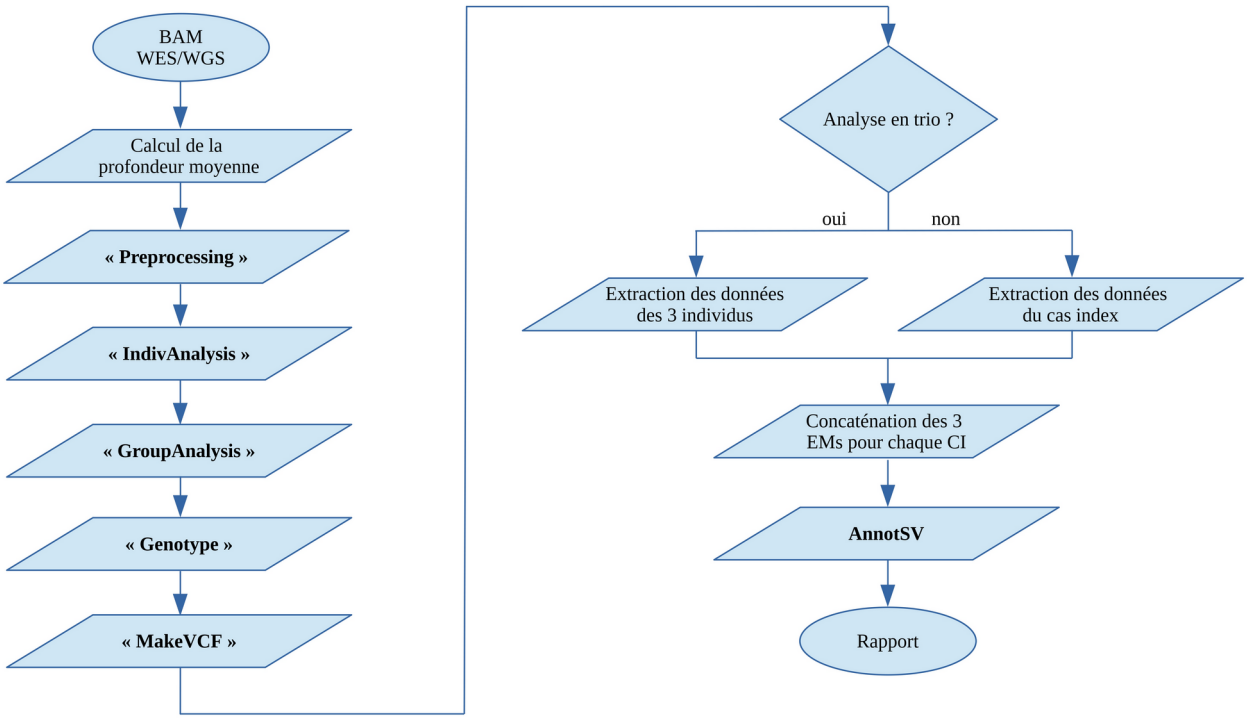


Figure 40 : Pipeline de détection des éléments mobiles avec l'outil MELT sur un groupe d'individus
L'analyse s'effectue par « batch » ou groupe d'individus. Les étapes « Preprocessing », « IndivAnalysis », « GroupAnalysis », « Genotype » et « MakeVCF » sont réalisées par le programme MELT. Des scripts python permettent le calcul de la profondeur, l'extraction des données des individus d'une même famille et l'obtention du rapport final pour chaque cas index.

Les éléments mobiles restants ont ensuite été analysés manuellement et individuellement pour mettre en évidence une concordance phénotypique entre la description OMIM et le dossier clinique du patient considéré.

II.4.2- Pipeline Tangram

Les données d'exome des 3322 individus ont été analysées par Tangram indépendamment (**Fig. 41a**). Tangram est un outil initialement conçu pour analyser des fichiers BAM générés par l'aligneur MOSAIK. Un module de Tangram a permis de convertir les

données des patients obtenus par BWA (étape « tangram_bam »).

Le nouveau fichier BAM nécessitait d'être trié par chromosome (étape « sort ») et de posséder dans son header les empreintes md5 des chromosomes du génome humain (étape « reheader » par Samtools). Une empreinte md5 correspond à une suite de caractères alphanumériques unique pour chaque fichier ou chaîne de caractères qu'elle permet d'identifier. Ainsi toute modification de contenu du fichier ou de la chaîne de caractères modifie son empreinte. Le programme a ensuite déterminé la distribution des longueurs de fragments contenus dans chaque fichier BAM (« tangram_scan »). Puis, à partir du BAM, Tangram a détecté et génotypé chaque élément mobile (« tangram_detect ») avant de les rapporter dans un fichier VCF brut. Un fichier par chromosome a ainsi été généré. Les 25 fichiers VCF ont ensuite été concaténés. Les résultats obtenus pour chaque individu (cas index ou apparentés) ont été comparés aux EM candidats identifiés avec le programme MELT.

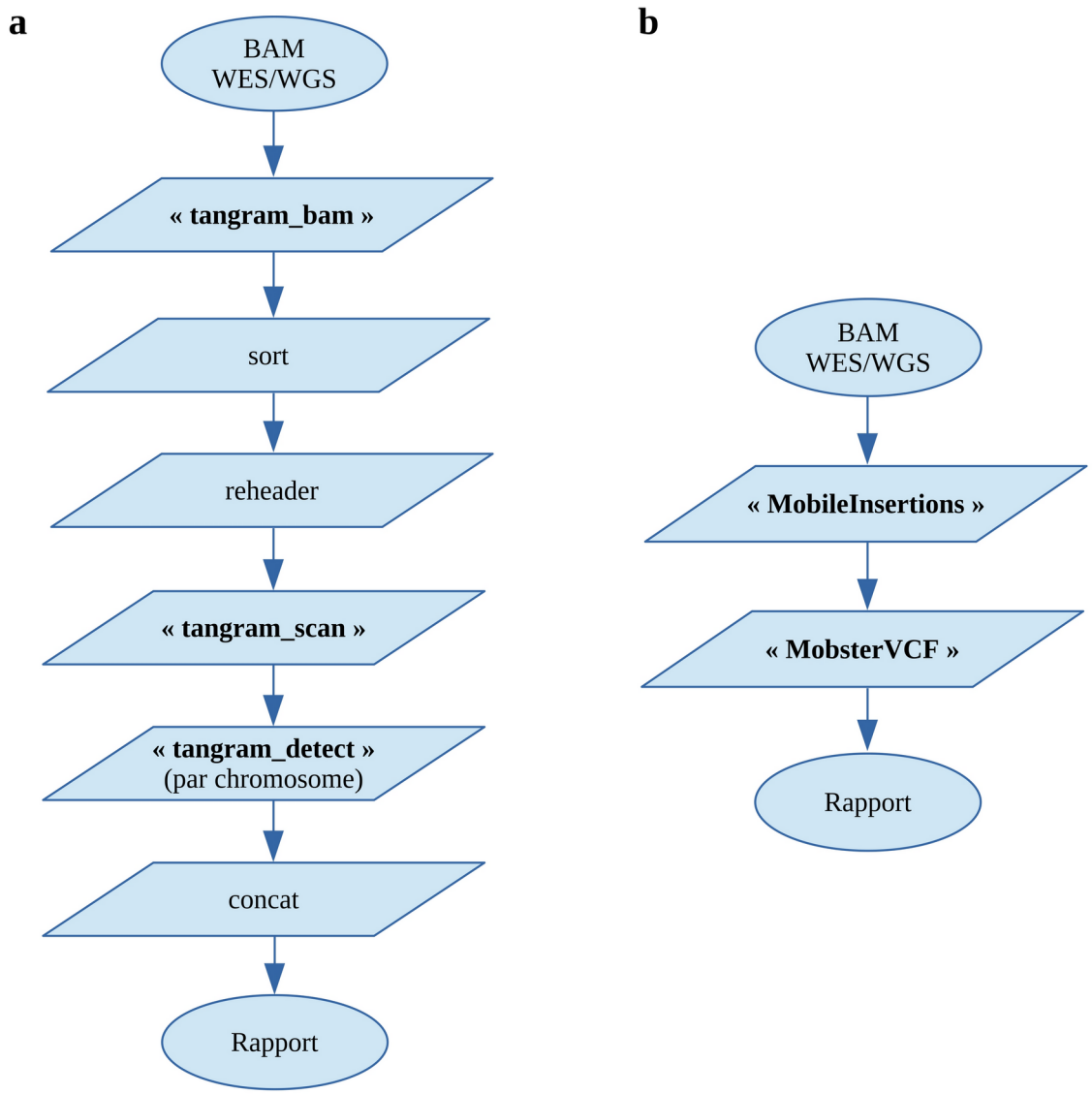


Figure 41 : Pipeline de détection des éléments mobiles avec les outils Tangram ou Mobster *a) Tangram accepte les données alignées par BWA grâce au module « tangram-bam » qui convertit les alignements BWA. Les fichiers BAM ainsi obtenus sont modifiés (« sort » et « reheader ») avec samtools afin que le programme puisse déterminer la distribution de la longueur de fragments (module « tangram_scan »). Tangram détecte ensuite les EM de chaque chromosome grâce au module « tangram_detect ». Un script concatène ensuite les 25 fichiers VCF obtenus. b) Le module « MobileInsertions » de Mobster réalise toutes les étapes nécessaires à la détection des EM (détection des paires discordantes et localisation de l'insertion). Le module « MobsterVCF » permet d'obtenir les résultats au format VCF.*

II.4.3- Pipeline Mobster

Les données des 3322 exomes ont été analysées indépendamment par Mobster (**Fig. 41b**). Il s'agit d'un programme qui accepte les données sous forme de fichiers BAM générés par MOSAIK ou BWA. Le pipeline comporte deux étapes principales qui consistent en l'analyse du fichier BAM par Mobster et en la conversion du fichier de sortie au format VCF. Le fichier de sortie du programme Mobster est un fichier texte tabulé de résultats bruts. L'analyse bioinformatique des données a consisté tout d'abord en l'identification des lectures indiquant l'insertion potentielle d'un EM (étape « PotentialMEIReadFinder » du module « MobileInsertions »). Puis ces lectures retenues ont été alignées sur le mobilome par MOSAIK. Puis Mobster a reconstitué les DP à partir des résultats des 2 étapes précédentes (étape « RefAndMEPairFinder » du module « MobileInsertions »). Enfin le programme a clusterisé les lectures alignées en 5' et en 3' de l'élément mobile suspecté afin de prédire la nature de ce dernier, la région et/ou le point d'insertion au sein du génome de référence (étape « AnchorClusterer » du module « MobileInsertions »). Mobster a ainsi généré le fichier de résultats qui a été converti en VCF par un autre module du programme (« MobsterVCF »). Les résultats obtenus pour chaque individu (cas index ou apparentés) ont été comparés aux EM candidats identifiés avec le programme MELT.

II.5- Validation des éléments mobiles candidats par séquençage ciblé d'ADN

PCR et séquençage MiSeq des éléments mobiles candidats : La PCR a été utilisée pour vérifier l'existence d'insertions des éléments mobiles potentiels au niveau des points de cassure suspectés. Les régions d'intérêt sont amplifiées avec le kit PrimeStar GXL (Takara Bio Inc.) selon les instructions du fournisseur. Le temps d'élongation varie en fonction de la taille du fragment attendue : 2 min pour une insertion d'un Alu et 7 min pour une insertion d'un LINE-1. La région d'intérêt a été amplifiée par PCR avec amorces spécifiques (**Tableau 8**) : une étape

de 3 minutes à 98°C, 5 cycles de 30 secondes à 94°C suivie de 30 secondes à 65°C et du temps d'élongation à 68°C, puis 35 cycles de 30 secondes à 94°C suivie de 30 secondes à 60°C et du temps d'élongation à 68°C et une étape finale de 10 minutes à 72°C. L'amplification par PCR est ensuite vérifiée par électrophorèse sur gel d'agarose en TBE à 1 %.

Gène	**Primers PCR**		**Temps d'élongation**
	Forward (5' → 3')	**Reverse (5' → 3')**	
ADGRG6	TCAGCCAAGAGTGTTTTCATATCAA	TAACTTGTCTGTGCATCCCTTTAAA	2 min
NPRL3	CAACAAGAGCTTTGACCAGAGG	ACTGATGATTCTCCCCAGCAT	2 min
FERMT1	TAGTCATGTTACAACTGGCTCTTTC	GTATTCCTTTCACTTACCTGAGCTG	2 min
SLC26A2	CTGGTATTTTCTCTGGTGTAGGAAG	GCATAGCAACTTTTGTCACATATCC	2 min
KMT2D	CTCTTGGCCCTTTCAATGAGTC	TACATTGCCACTCAGTTACCCTTAA	2 min
GRIN2B	TTCTTCTTCTGGGCCTTGGAATTAG	GTAAGCAGCAATATAAGGACAGCC	2 min
SETD5	CAGAGTTCAACTTGATGTATGCCTA	TTGTACTAAAACAGCCTGTCTTTCA	7 min
TTN	GTGCTTCTTGGCTCATTCTTATTTC	ACAGGTCTTCGTCGGATTATAAAAG	7 min
SYNE1	TTTTCAATGTTCTGTCGGTGATTTG	AATTTAGTGGCATGTGAGACAGTC	7 min

Tableau 8 : Conditions des analyses par PCR des éléments mobiles candidats

Extraction d'ADN amplifié sur gel TAE : Les produits de PCR ont été mis à migrer sur gel TAE (Tris/Acétate/EDTA) à 1,5 % d'agarose. Après découpe des bandes d'intérêt, l'ADN a été extrait grâce au kit MinElute Gel Extraction de Qiagen selon les instructions du fournisseur.

Séquençage par MiSeq : Les amplicons d'un échantillon sont purifiés à l'aide de billes magnétiques AMpure XP (Beckman Coulter Inc.). Ces banques d'amplicons sont utilisées pour préparer les librairies de séquençage à l'aide du kit Nextera XT (Illumina Inc.) et selon les instructions du fournisseur. Le séquençage est ensuite réalisé sur le séquenceur haut débit MiSeq (Illumina Inc.) en lectures pairées de 150 pb.

II.6- Validation des éléments mobiles candidats par étude de leurs impacts sur l'épissage et l'expression géniques

Extraction, dosage et contrôle qualité des ARN totaux du sang : Les ARN totaux sont extraits à partir de sang total prélevé en tube PAXgene (Preanalytics GmbH) à l'aide du kit PAXgene Blood RNA (Preanalytics GmbH) suivant les instructions du fournisseur. Les ARN totaux sont ensuite dosés par mesure de l'absorbance (Multiscan GO, ThermoFischer Scientific Inc.), leur qualité est évaluée en déterminant le RNA integrity number (RIN) à l'aide du kit RNA nano 6000 et du bioanalyzer 2100 (Agilent Technologies Inc.).

Extraction et dosage des ARN totaux de fibroblastes : Les ARN totaux des cultures de fibroblastes ont été extraits avec du TRIzolReagent de Life Technologies selon les instructions du fournisseur. La concentration a été déterminée avec le NanoDrop (Thermo Fisher Scientific).

Amorces pour l'analyse de l'ARN :

Nom de l'amorce	Primers		Fonctions
	Forward (5' → 3')	Reverse (5' → 3')	
FERMT1_cDNA_ex5-10F ou R	AAACCATGACCCCTATATATGACCC	GGTTCTCCTTGTTCAAGTTCCTTAT	Étude de l'épissage
FERMT1_qPCR_ex2F ou R	CACTGACTTTACATTTGCTTCCTG	GATACTCTCAGTGTGACGTCTTTC	qPCR
FERMT1_qPCR_ex12F ou R	GTCCTCAACATCCTTTCATTTCTGA	ACAAAACATTCTGGGTTCATATCCA	qPCR

Tableau 9 : Amorces pour les analyses de l'ARN du gène FERMT1

Étude de l'impact d'une variation sur l'épissage (séquençage de l'ADNc) : L'ADNc a été obtenu à partir d'ARN grâce au kit QuantiTect Reverse Transcription (Qiagen GmbH) et selon les instructions du fournisseur. L'ensemble des séquences codantes des régions d'intérêt ont été extraites à partir du site UCSC genome browser (http://genome-euro.ucsc.edu/index.html). Les amorces (**Tableau 9**) ont été positionnées dans les exons encadrant l'exon contenant la variation

d'intérêt à l'aide du programme Primer3 (Untergasser et al., 2012). L'amplification par PCR et le séquençage ont été réalisés comme décrit précédemment. Les données de séquençage ont été alignés avec STAR (v2.5.2b) (Dobin et al., 2013).

Étude de l'impact d'une variation sur l'expression génique (RT-qPCR) : Les couples d'amorces (**Tableau 9**) ont été positionnés dans les exons encadrant l'exon contenant la variation d'intérêt à l'aide du programme Primer3 (Untergasser et al., 2012). La taille des amplicons devait être comprise entre 90 et 110 pb. L'efficacité de l'amplification a été validée sur une gamme étalon d'ADN contrôle (32, 8, 2, 0.5, 0.125 et 0.03125 ng/µL). Pour chaque puits 5 µL de SsoAdvanced Universal SYBR Green Supermix de Bio-Rad (à une concentration de 2X), 0,2 µL par amorce (à une concentration de 10 µM), 2,6 µL d'eau qualité biologie moléculaire et 2 µL d'ADNc ont été déposés. La qPCR a été réalisée sur un thermocycleur avec bloc optique (CFX96 ou CFX384 de Bio-Rad) selon les instructions du fournisseur du SsoAdvanced Universal SYBR Green Supermix. Les résultats ont été analysés avec le logiciel CFX Manager de Bio-Rad.

Les ADNc témoins et l'ADNc de l'échantillon ont été dilués à 10 ng/µL. Pour les deux couples d'amorces (**Tableau 9**) et les couples contrôles (*TRPV4* et *VPS41*), l'échantillon, l'ADNc contrôle et le contrôle négatif (eau) ont été déposés en triplicats. La qPCR a été réalisée comme décrit précédemment.

III- RÉSULTATS : MISE EN ÉVIDENCE D'ÉLÉMENTS MOBILES AU SEIN DE GÈNES OMIM MORBIDES

III.1- Analyses des résultats bioinformatiques

L'étude des contrôles positifs analysés par MELT, Tangram et Mobster a montré que ces pipelines permettaient bien de mettre en évidence des EM (**Fig. 42**). Le pipeline MELT

donnait les meilleurs résultats : chaque famille d'EM a été retrouvées avec des valeurs cohérentes. Le coefficient de corrélation de Pearson était de 0,97 entre les deux jeux de données (**Tableau 10**). Les coefficients de Pearson pour les outils Tangram et Mobster mettaient en évidence une plus grande différence entre les résultats attendus (contrôles) et ceux de ces 2 outils.

Contrôles vs	Total des EM	Alu	L1	SVA
MELT	0,9705	0,9672	0,9471	0,9471
Tangram	0,4384	0,5816	0,3050	0,4266
Mobster	0,7336	0,7848	0,6699	0,8980

Tableau 10 : Comparaison des résultats des 3 outils (coefficients de corrélation de Pearson)

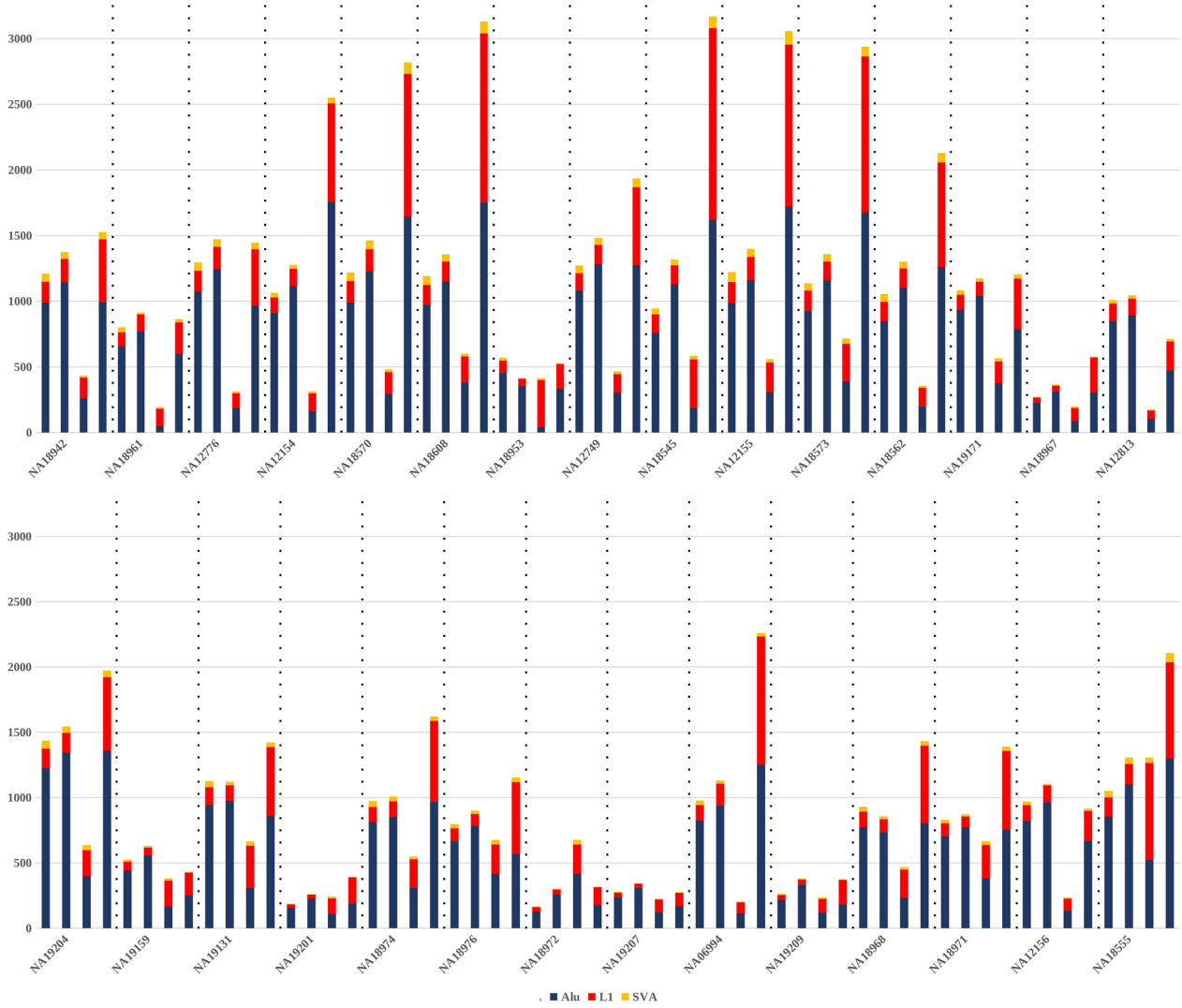


Figure 42 : Nombre d'éléments mobiles identifiés par le consortium 1000 Génomes comparés aux pipelines MELT, Tangram et Mobster chez 29 individus de la cohorte 1000 Génomes *Pour chaque individu, les valeurs du consortium sont présentées à gauche, suivies de celles des pipelines MELT, Tangram et Mobster. Les effectifs des Alu sont indiqués en bleu, des L1 en rouge et des SVA en jaune.*

Pour 3322 individus (dont 2500 cas index) analysés avec l'outil MELT, 2394 cas index ont eu un résultat brut non négatif. 496 312 éléments mobiles potentiels ont été détectés avec 404 236 (81,45 %) Alu, 87751 (17,68 %) L1 et 4325 (0,87 %) SVA (**Fig. 43**). En ne conservant que les EM insérés au sein d'un gène OMIM morbides, 84 % des éléments détectés ont été éliminés. Parmi les 79 392 candidats potentiels, 4884 (93,84 %) étaient présents au sein de régions non introniques (exon, promoteur, terminateur, 5' et 3'-UTR). Les éléments présents à

une fréquence supérieure à 1 % dans la base de données 1000 Génomes n'ont également pas été retenus dans la liste des candidats (-13,23 % des effectifs déjà filtrés). Les résultats fournis par le pipeline indiquent le nombre d'observations dans le batch de chaque EM potentiel détecté. N'ont été conservés que les éléments présents moins de 5 fois dans le batch considéré. Ainsi, toute insertion homozygote héritée des 2 parents n'aura donc pas été éliminée sur ce critère. De même, une éventuelle insertion transmise par les 2 parents à leurs 2 enfants (unique famille de 4 individus de la cohorte) n'aura pas été filtrée sur cette fréquence. Les effectifs déjà filtrés ont alors été réduit de 73,67 %. Après concaténation des résultats restants, les éléments mobiles observés plus de 5 fois dans la cohorte des cas index ont été éliminés. Les 1070 éléments mobiles potentiels à l'issue des étapes de filtres ont été identifiés chez 516 cas index. Ainsi les filtres ont permis de réduire le nombre de candidats de plus de 99,78 % (n=495 242). Les candidats restants sont à 50,47 % (540/1070) des Alu, 49,25 % (527/1070) des L1 et 0,28 % (3/1070) des SVA (**Fig. 43**). La moyenne est ainsi de 2,07 EM par patient restant après les filtres avec une médiane à 1. Les 1070 éléments mobiles obtenus ont été ensuite filtrés par concordance phénotypique entre les patients concernés et la littérature.

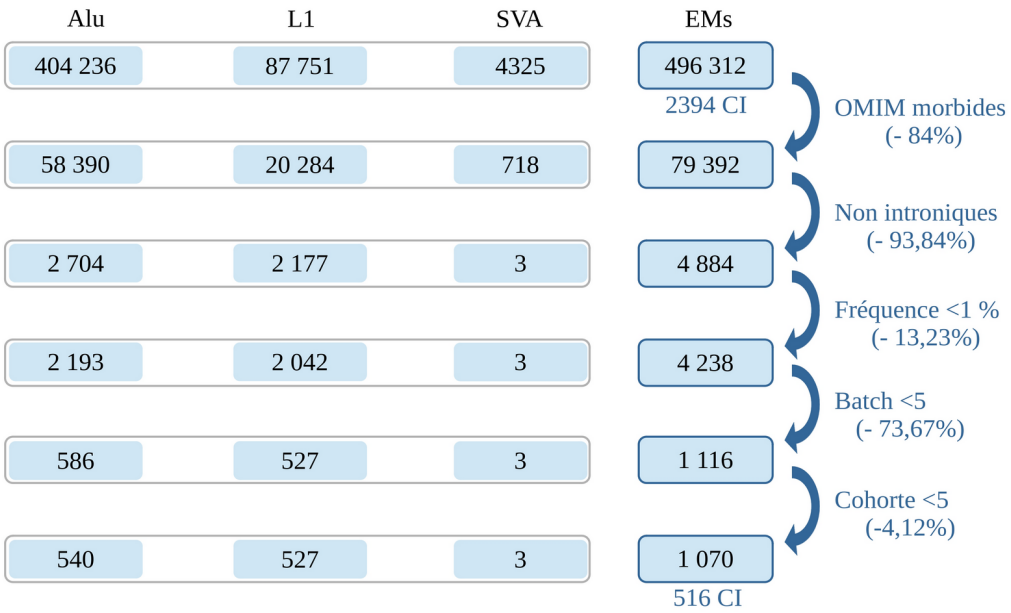


Figure 43 : Filtres appliqués aux EM détectés des 2394 cas index ayant un résultat suite à l'analyse par MELT

La majorité des éléments détectés sont des Alu, suivis des LINE-1 puis des SVA. Les 5 filtres appliqués et indiqués à droite ont permis de réduire de plus de 99 % le nombre de candidats potentiels. Environ 78 % des patients ont ainsi été retirés de la liste.

A l'issue de l'analyse des concordances phénotypiques, 9 éléments mobiles candidats ont été retenus au sein des gènes *ADGRG6*, *NPRL3*, *FERMT1*, *SLC26A2*, *KMT2D*, *SETD5*, *TTN*, *SYNE1* et *GRIN2B* (**Tableau 11**). Trois sont des éléments LINE-1 (*SETD5*, *TTN* et *SYNE1*) et 6 des éléments Alu (*ADGRG6*, *NPRL3*, *FERMT1*, *SLC26A2*, *KMT2D* et *GRIN2B*).

Gènes (EM)	MELT		Tangram	
	Position	GT	Position	GT
ADGRG6 cas index (Alu)	chr6:142741104	0/1	x	x
NPRL3 cas index (Alu)	chr16:135837	0/1	chr16:135837	0/1
NPRL3 père (Alu)	chr16:135837	0/1	x	x
FERMT1 cas index (Alu)	chr20:6078235	0/1	x	x
SLC26A2 cas index (Alu)	chr5:149357663	0/1	x	x
KMT2D cas index (Alu)	chr12:49438216	0/1	x	x
GRIN2B cas index (Alu)	chr12:13716543	0/1	x	x
SETD5 cas index (LINE-1)	chr3:9512598	0/1	x	x
TTN cas index (LINE-1)	chr2:179616148	0/1	x	x
SYNE1 cas index (LINE-1)	chr6 :152642851	0/1	x	x

GT : Génotype ; NA : non disponible

Tableau 11 : Éléments mobiles candidats détectés par MELT et résultats obtenus par Tangram et Mobster

III.2- Validation des EM candidats identifiés *in silico*

Les candidats insérés au sein des gènes *ADGRG6*, *NPRL3*, *FERMT1*, *SLC26A2*, *KMT2D*, *SETD5*, *TTN* et *SYNE1* ont d'abord été vérifiés par PCR (**Fig. 44**). Selon les résultats de MELT une insertion hétérozygote est attendue pour chacun de ces 8 candidats. La confirmation et la ségrégation de l'insertion du gène *GRIN2B* ont été effectuées en simultané (**Fig. 50**).

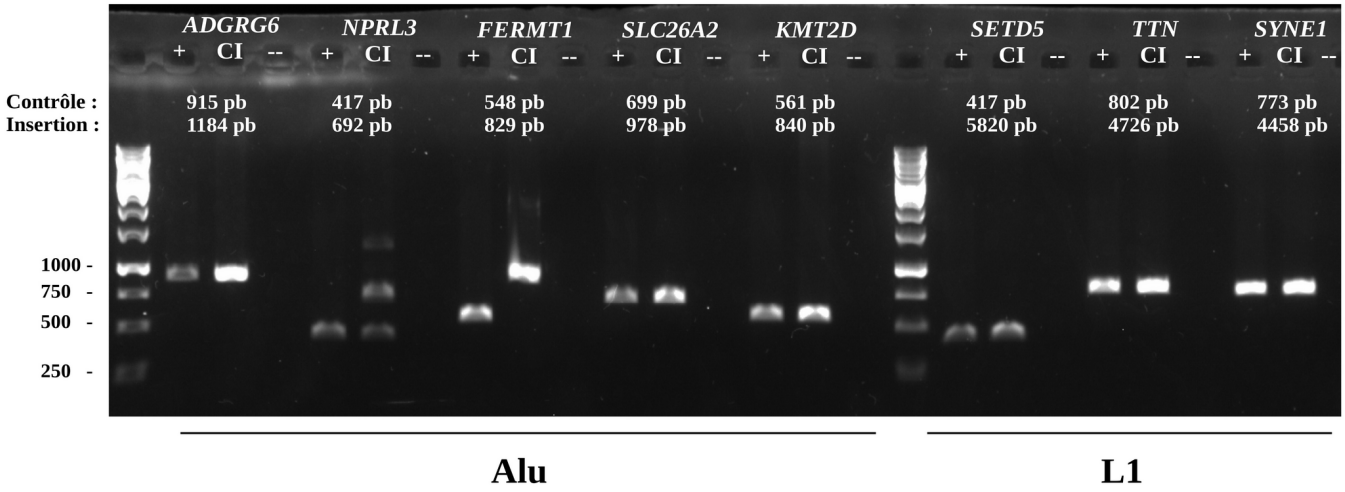


Figure 44 : PCR des éléments mobiles candidats dans les gènes ADGRG6, NPRL3, FERMT1, SLC26A2, KMT2D, SETD5, TTN* et *SYNE1

La partie gauche du gel concerne les insertions d'Alu, la partie droite les insertions de L1. Tailles de fragments attendues en paires de bases (pb) sans insertion/avec insertion pour ADGRG6 : 915/~1184, NPRL3 : 417/~692, FERMT1 : 548/~829, SLC26A2 : 699/~978, KMT2D : 561/~840, SETD5 : 417/~5820, TTN : 802/~4726 et SYNE1 : 773/~4458. (+) contrôle de l'amplification du fragment sans insertion. CI : cas index. (–) contrôle négatif de la PCR.

Aucune amplification de fragments de taille attendue en cas d'insertion n'a été observée pour les gènes *ADGRG6*, *SLC26A2*, *KMT2D*, *SETD5*, *TTN* et *SYNE1* (**Fig. 44**). Il s'agit donc de faux positifs. Les profils des gènes *FERMT1*, *GRIN2B* et *NPRL3* ont présenté des bandes correspondant à des insertions. La ségrégation des insertions dans les gènes *FERMT1*, *GRIN2B* et *NPRL3* a ensuite été réalisée par PCR.

Étude du gène *FERMT1*

Le patient est un homme âgé de 80 ans adressé pour leucokératose buccale avec poïkilodermie acquise. Il présente une microstomie (**Fig. 45a**), une leucokératose labiale, palatine et jugale droite, une poïkilodermie depuis l'enfance (cou, aisselles, avant-bras) (**Figs. 45c et 45d**), une maladie de Dupuytren bilatérale, une kératodermie palmaire (**Fig. 45e**) et une dystrophie unguéale avec trachyonychie. L'exome solo est revenu non concluant. Les hypothèses diagnostiques formulées par le clinicien étaient : une dyskératose congénitale de Zinsser-Cole-Engman de forme atypique, une poïkilodermie bulleuse de Tessa-Kindler ou une poïkilodermie de forme atypique.

Figure 45 : Photos du patient 1 atteint de poïkilodermie
a) Limitation de l'ouverture buccale et chéilite (inflammation). b) Ectropion bilatéral (affaissement de la paupière). c) et d) poïkilodermie. e) Kératodermie palmaire. (Consentement du patient pour la diffusion de ses photos).

Le gène *FERMT1* (OMIM 173650) est impliqué dans le syndrome de Kindler de transmission autosomique récessive, donnant une très forte corrélation entre les informations cliniques et biologiques.

La migration sur gel d'agarose du produit de PCR a bien confirmé l'insertion d'un fragment de taille similaire à celle d'un élément Alu (**Fig. 46**). Seule une bande d'environ 829 pb été retrouvée chez le cas index quel que soit le tissu prélevé. Contrairement aux informations fournies par les résultats de MELT, il s'agirait d'une insertion homozygote. Le profil IGV a d'ailleurs confirmé cette hypothèse. L'ADN des parents du patient n'étant pas accessible, le profil PCR de ces derniers n'a pas pu être réalisé. En revanche la ségrégation chez les 3 enfants du patient a montré que ces derniers étaient hétérozygotes pour l'insertion : bande à ~548 pb (bande 3) et à ~829 pb (bande 2). De plus, il existe une autre bande (bande 1) qui semble de taille plus importante (~2500 pb) et qui correspondrait à la formation d'hétéroduplex. L'hybridation d'un allèle muté avec un allèle WT provoque la modification de la structure 3D du duplex à l'origine d'un ralentissement de la migration (**Fig. 46**). La recherche de ce type de conformation servait de pré-criblage de variations de séquence (substitutions, indels) avant séquençage par Sanger. Plusieurs méthodes de mise en évidence des hétéroduplex existent (Balogh et al., 2004) et font appel à la migration d'ADN double-brin. Les conditions de la méthode CSGE (Conformation-Sensitive Gel Electrophoresis) se rapprochent de celles utilisées pour la ségrégation des éléments mobiles (pas de gradient de température ou d'agent dénaturant).

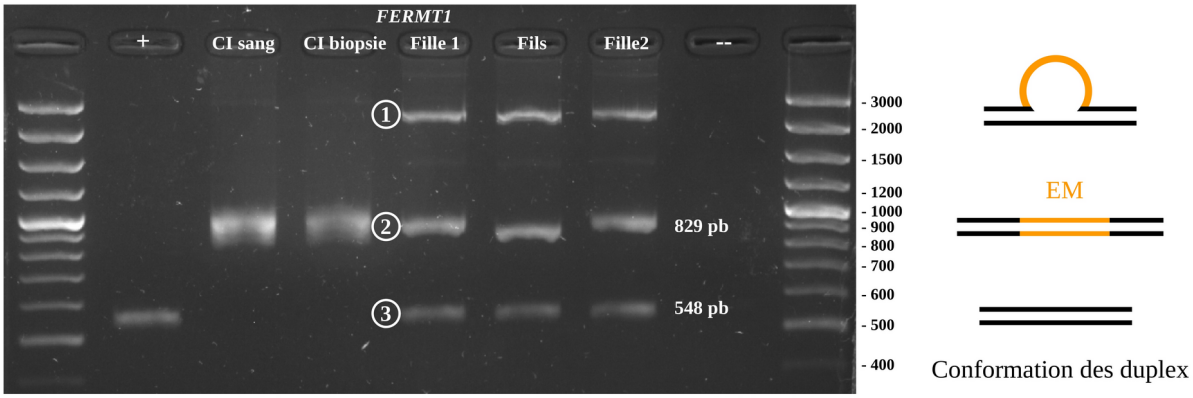


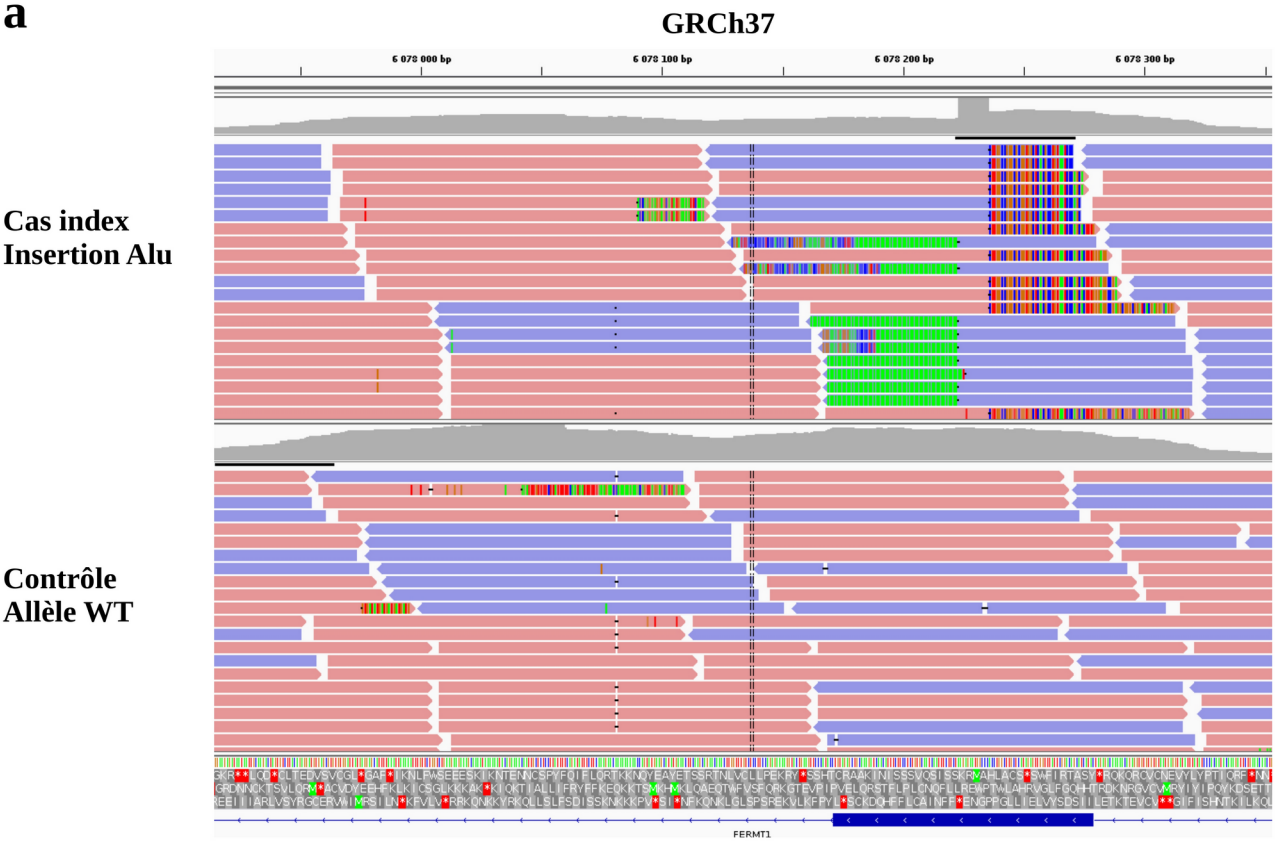
Figure 46 : Validation et ségrégation de l'élément mobile candidat du gène FERMT1 par migration du produit de PCR sur gel d'agarose

Ségrégation de l'insertion chez le cas index (CI) (sang + fibroblastes) et ses enfants. Tailles de fragments attendues en paires de bases (pb) sans/avec insertion : 548/~829. Conformation des duplex en fonction de leur migration. (+) contrôle de l'amplification du fragment sans insertion. (–) contrôle négatif de la PCR.

Afin de vérifier la nature des fragments insérés, un séquençage MiSeq de la bande du cas index, après extraction sur gel, a été réalisé.

Les séquences obtenues par MiSeq ont été alignées sur le génome de référence, la séquence du gène *FERMT1*, la séquence du gène avec insertion et sur la séquence de l'élément Alu. L'alignement sur le génome de référence des fragments pour le gène *FERMT1* montre uniquement la présence de lectures splittées chez le cas index au niveau du point de cassure suspecté (**Fig. 47a***)*. Comme attendu la bande contrôle ne présente pas d'anomalie à l'alignement. L'alignement sur les séquences de référence des éléments mobiles (non présenté) permet d'identifier la présence de l'élément AluY au sein des fragments d'ADN du patient. Dans un second temps, les données ont été alignées sur la séquence du gène *FERMT1* modifiée par l'ajout des séquences Alu, polyA et TSD au niveau du point de cassure détecté *in silico* (**Fig. 47b**). Sur le profil du cas index, on n'observe pas de lectures splittées qui indiqueraient la présence de fragments sans insertion. Il s'agit d'un alignement à couverture homogène, en dehors de la queue polyA. L'amplification de la région répétée de la queue polyA conduit à des répétitions de nucléotides A de longueurs différentes. L'alignement de ces fragments peut donc présenter des insertions ou des délétions de bases A. Le profil du contrôle présente une interruption dans l'alignement des lectures au niveau de l'insertion de l'élément Alu. Les lectures « splittées » ne s'alignant que sur la référence sont présentes. Des lectures alignées de façon non spécifique sont observées mais leur nombre est bien inférieure à celles spécifiques de la référence. Les résultats du MiSeq pour le gène *FERMT1* confirment donc la présence d'une insertion homozygote d'un élément identifié comme AluY chez le cas index.

a


GRCh37
6 078 000 bp
6 078 100 bp
6 078 200 bp
6 078 300 bp
Cas index
Insertion Alu
Contrôle
Allèle WT
FERMT1

b

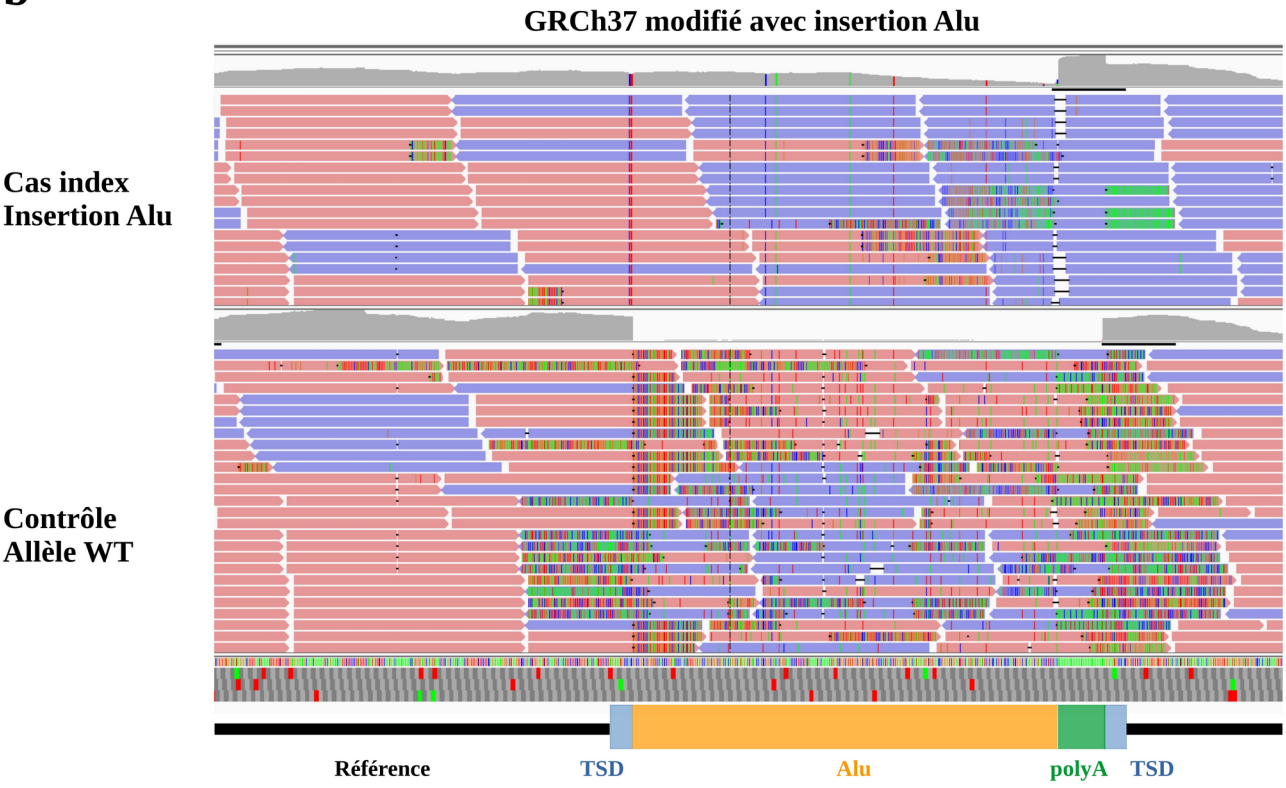

GRCh37 modifié avec insertion Alu
Cas index
Insertion Alu
Contrôle
Allèle WT
Référence
TSD
Alu
polyA
TSD

Figure 47 : Vue IGV de l'insertion de l'élément Alu + TSD au sein de l'exon 7 du gène FERMT1 chez le cas index et un individu contrôle (sans insertion)
L'intégralité du produit de PCR du cas index (bande 2) et d'un contrôle sans insertion (bande 3) a été séquencé. a) Alignement sur GRCh37. L'insertion d'un élément mobile chez le cas index est identifiée par les lectures « splittées ». Le profil du contrôle ne met pas en évidence d'insertion. b) Alignement sur FERMT1 avec insertion. Profil du cas index : Les données de séquençage s'alignent sur cette référence modifiée par insertion d'un élément Alu. La couverture de cette région est continue. Profil du contrôle : L'absence de séquence Alu au sein cet allèle WT est à l'origine de l'absence de lecture alignée au niveau de l'insertion. Seules des lectures « splittées » sont présentes.

Les produits de PCR des 3 enfants ont également été séquencés (**Figs. 48 et 49**). On observe chez la fille aînée (fille 1) (**Fig. 48**) la même présentation que pour le cas index du gène *NPRL3* au niveau des 3 profils de bandes pour l'alignement sur les deux références. Il existe bien chez la fille aînée l'allèle WT et l'allèle porteur de l'insertion Alu. Les deux autres membres de la fratrie présentent les mêmes profils que la sœur aînée pour les deux alignements (**Fig. 49**). Ils sont donc bien porteurs d'une insertion Alu dans le gène *FERMT1* héritée du père (cas index). Les profils PCR et de MiSeq des 3 membres de la fratrie sont identiques : ils sont hétérozygotes pour l'insertion de l'élément Alu.

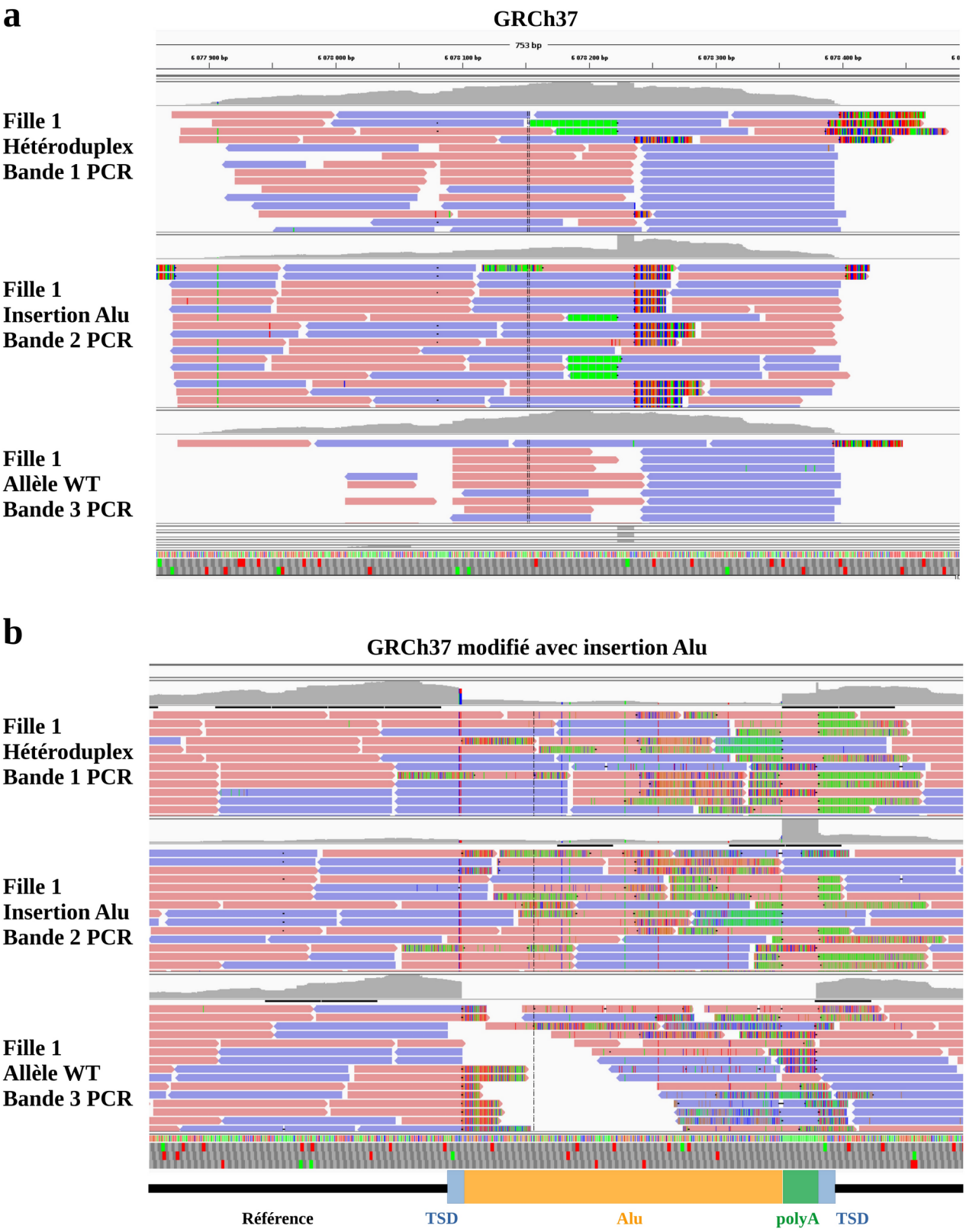
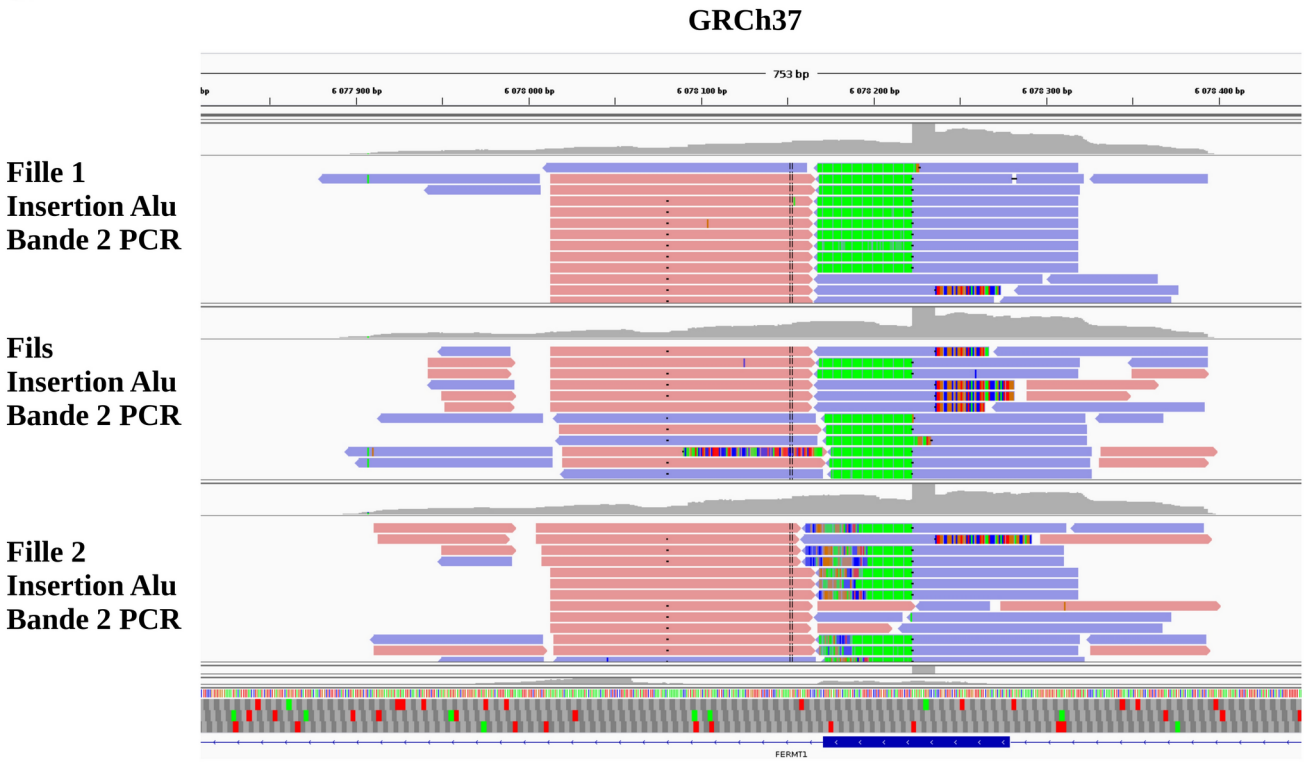

a
GRCh37
753 bp
Fille 1
Hétéroduplex
Bande 1 PCR
Fille 1
Insertion Alu
Bande 2 PCR
Fille 1
Allèle WT
Bande 3 PCR
b
GRCh37 modifié avec insertion Alu
Fille 1
Hétéroduplex
Bande 1 PCR
Fille 1
Insertion Alu
Bande 2 PCR
Fille 1
Allèle WT
Bande 3 PCR
Référence
TSD
Alu
polyA
TSD

Figure 48 : Vue IGV de l’insertion de l’élément Alu + TSD au sein de l’exon 7 du gène FERMT1 chez la fille aînée du cas index

Les 3 produits de PCR obtenus pour la fille aînée ont été séquencés indépendamment. Les profils sont présentés selon l'ordre de migration (bandes 1, 2 et 3 de la PCR). Les bases non-appariées sont représentées en couleur. a) Alignement sur GRCh37. La région du point de cassure est mise en évidence par les lectures « splittées » qui ne s’alignent que partiellement sur le génome de référence. Profil de la bande 1 : la coexistence de lectures « splittées » et de lectures normales renforce l’hypothèse de l’existence d’hétéroduplex allèle muté/allèle WT. Profil de la bande 2 : l’insertion d’un élément mobile est identifiée par les lectures « splittées ». Profil de la bande 3 : aucune anomalie d’alignement n’est identifiée pour la séquence de l’allèle WT. b) Alignement sur FERMT1 avec insertion. Profil de la bande 1 : L'alignement des données montre la présence de l’allèle WT et de l’allèle contenant l’insertion Alu, les TSD et sa queue polyA. Il existe des lectures "splittées" qui ne s'alignent que sur le génome de référence et qui présentent des mésappariements avec la région insérée. Ces lectures sont retrouvées sur le profil de la bande 3. Il existe en parallèle des lectures qui s'alignent sur la référence modifiée malgré la présence de l'insertion. Il y a donc deux fois moins de lectures qui couvrent la région insérée. Profil de la bande 2 : Les données de séquençage s’alignent sur cette référence modifiée par insertion d’un élément Alu. Profil de la bande 3 : L’absence de séquence Alu au sein cet allèle WT est à l’origine de l’absence de lecture alignée au niveau de l’insertion. Seules des lectures « splittées » sont présentes.

a


GRCh37
753 bp
Fille 1
Insertion Alu
Bande 2 PCR
Fils
Insertion Alu
Bande 2 PCR
Fille 2
Insertion Alu
Bande 2 PCR
FERMT1

b

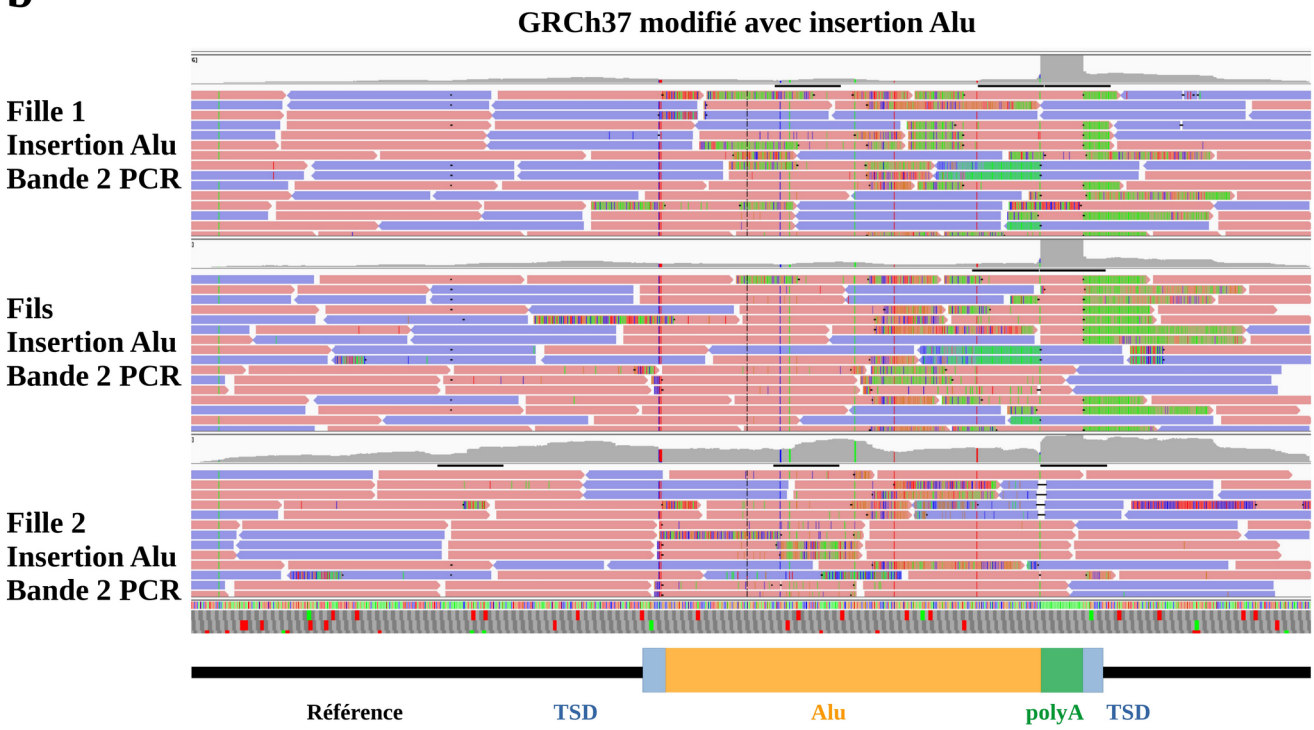

GRCh37 modifié avec insertion Alu
Fille 1
Insertion Alu
Bande 2 PCR
Fils
Insertion Alu
Bande 2 PCR
Fille 2
Insertion Alu
Bande 2 PCR
Référence
TSD
Alu
polyA
TSD

Figure 49 : Vue IGV de l'insertion de l'élément Alu + TSD au sein de l'exon 7 du gène FERMT1 chez les 3 enfants du cas index
Le produit de PCR correspondant à l'insertion de l'élément Alu (bande 2) a été séquencé chez les 3 enfants du cas index. Les bases non-appariées sont représentées en couleur. a) Alignement sur GRCh37. La région du point de cassure est mise en évidence par les lectures « splittées » qui ne s'alignent que partiellement sur le génome de référence. L'insertion d'un élément mobile est identifiée par les lectures « splittées ». b) Alignement sur FERMT1 avec insertion. Pour les 3 membres de la fratrie les données de séquençage s'alignent sur cette référence modifiée par insertion d'un élément Alu.

La ségrégation familiale et les données de MiSeq de l'insertion d'un élément Alu au sein de l'exon 7 du gène *FERMT1* avec l'individu homozygote atteint et les individus hétérozygotes non atteints, ainsi qu'un phénotype clinique tout à fait spécifique plaide en faveur de la pathogénicité de cette insertion Alu. Une analyse de l'expression et de l'épissage de l'ARN des fibroblastes du patient est en cours pour conforter ces résultats.

Étude du gène GRIN2B

Il s'agit d'une patiente née à terme en 2015 (5 ans) adressée pour déficience intellectuelle légère. Cette patient, séquencée en solo, présente une hypotonie axiale, un retard du développement, des stéréotypies gestuelles et une déficience intellectuelle. Aucune particularité morphologique n'a été rapportée. Ses mensurations à la naissance étaient dans la norme. Les parents ne sont pas atteints de la pathologie de leur fille. La lecture diagnostique de l'exome en solo est revenue négative.
Le gène *GRIN2B* (OMIM 138252) a été décrit dans l'encéphalopathie épileptique infantile précoce de transmission autosomique dominante et dans la DI de sévérité variable également de transmission autosomique dominante. Les variations décrites sont majoritairement *de novo*. Ainsi, une insertion *de novo* de l'élément mobile chez le cas index renforcerait l'hypothèse d'un impact de cette rétrotransposition sur le phénotype de la patiente.

La migration sur gel d'agarose du produit de PCR a bien confirmé l'insertion d'un fragment de taille similaire à celle d'un élément Alu chez le cas index (**Fig. 50**). Comme attendu l'insertion est hétérozygote car la bande contrôle (640 pb, bande 3) et la bande

d'insertion (~915 pb, bande 2) ont été retrouvées. La bande qui semble avoir une taille d'environ 3000 pb (bande 1) mettrait en évidence la formation d'hétéroduplex. Les profils parentaux correspondent à ceux du contrôle sans insertion et ne montrent pas d'insertion au sein de cette région génique. Afin de vérifier la nature des fragments insérés, un séquençage MiSeq des 3 bandes, après extraction sur gel, est en cours chez les 3 membres de la famille.

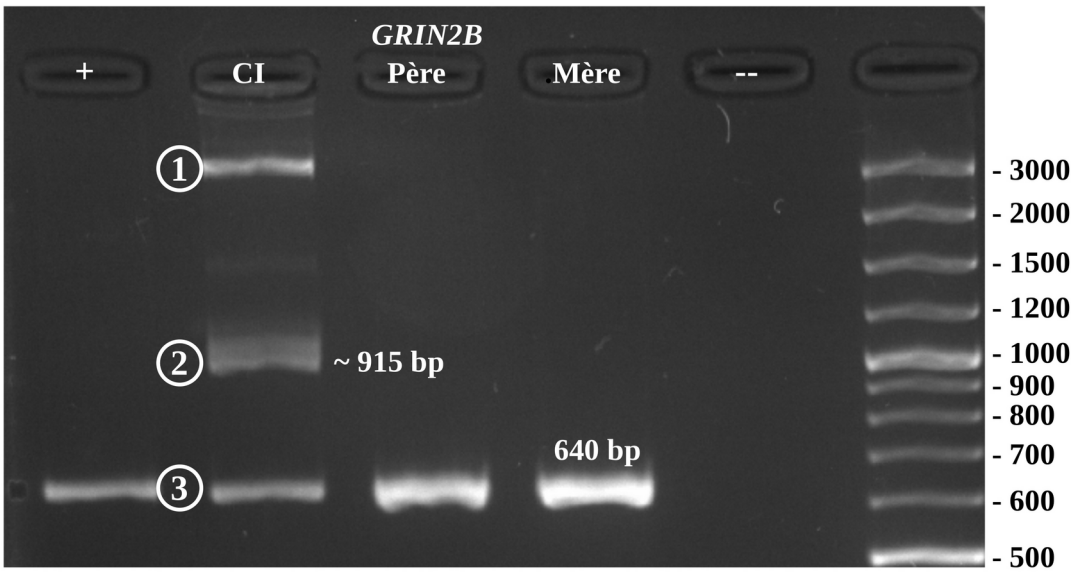


Figure 50 : Validation et ségrégation de l'élément mobile candidat au sein du gène GRIN2B par migration du produit de PCR sur gel d'agarose
Tailles de fragments attendues en paires de bases (pb) sans/avec insertion pour GRIN2B : 640/~915. CI : cas index. (+) contrôle de l'amplification du fragment sans insertion. (–) contrôle négatif de la PCR.

La ségrégation familiale et le phénotype clinique sont en faveur de l'implication de cette insertion dans la pathologie de la patiente. Une analyse de l'ARN sera donc réalisée afin de conforter ces résultats.

Étude du gène NPRL3

Il s'agit d'un garçon né en 2008 (11 ans) adressé pour une DI syndromique. Ce patient, séquencé avec ses parents (trio), présente un retard du développement global modéré, une DI modérée, des interactions sociales altérées, des convulsions généralisées tonico-cloniques, une hypoplasie du corps calleux et une fibromatose gingivale (hyperplasie gingivale). Son poids et sa taille à la naissance étaient dans la moyenne, son périmètre crânien à -2,3 DS. À 10 ans, ses mensurations étaient dans la moyenne. La lecture diagnostique de l'exome en trio est revenue négative et la lecture recherche a mis en évidence un gène candidat.
Le gène *NPRL3* (OMIM 617118) a été décrit dans l'épilepsie focale familiale à foyers variables de transmission autosomique dominante à pénétrance incomplète et qui se manifeste dans la petite enfance.

La migration sur gel d'agarose du produit de PCR a bien confirmé l'insertion d'un fragment de taille similaire à celle d'un élément Alu (**Fig. 51**). L'insertion est hétérozygote car la bande contrôle (417 pb, bande 3) et la bande d'insertion (~750 pb, bande 2) ont été retrouvées (**Fig. 51**). La dernière bande qui semble avoir une taille d'environ 2500 pb (bande 1) correspondrait à la formation d'hétéroduplex. Le profil du patient est, comme attendu selon les analyses de MELT, identique à celui du père (**Fig. 51**). Le profil de la mère est identique à celui du contrôle sans insertion. Afin de vérifier la nature des fragments insérés, un séquençage MiSeq des 3 bandes, après extraction sur gel, a été réalisé chez les 3 membres de la famille.

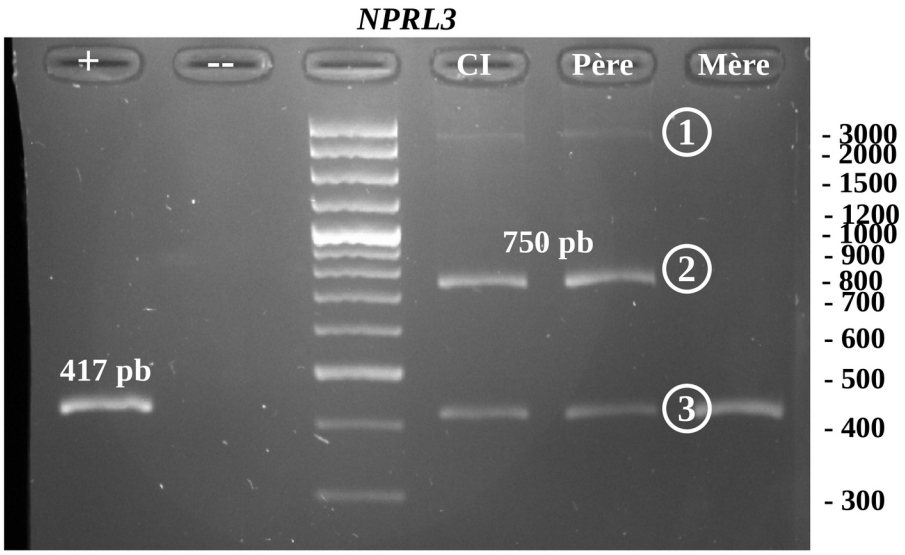


Figure 51 : Validation et ségrégation de l'élément mobile candidat du gène NPRL3 par migration du produit de PCR sur gel d'agarose
Confirmation de la ségrégation familiale de l'insertion au sein du gène NPRL3. Tailles de fragments attendues en paires de bases (pb) sans/avec insertion : 417/~692. CI : cas index. (+) contrôle de l'amplification du fragment sans insertion. (–) contrôle négatif de la PCR.

Les séquences obtenues par MiSeq ont été alignées sur le génome de référence, la séquence du gène *NPRL3*, la séquence du gène avec insertion et sur la séquence de l'élément Alu. L'alignement sur le génome de référence des fragments du cas index et du père (non présenté) pour le gène *NPRL3* confirme l'insertion hétérozygote d'un fragment d'ADN (**Fig. 52a**). La bande 1 présente des lectures « splittées » (témoin de la présence d'un élément mobile) et des lectures normales au niveau du point de cassure. La bande 2 ne contient que des fragments d'ADN avec insertion d'une séquence, identifiée par la présence unique de lecture « splittées » au niveau du point de cassure. La bande 3, de taille plus petite, ne contient que des lectures normales comme attendu en absence d'insertion. Ces données ont été alignées sur la séquence AluY (non présenté). Les résultats confirment la présence d'un élément AluY au sein de ces fragments d'ADN.

Puis les données ont été alignées sur la séquence du gène *NPRL3* modifié par l'ajout de la séquence AluY et polyA au niveau du point de cassure détecté i*n silico* (**Fig. 52b**). Le TSD n'ayant pas été détecté par le programme MELT, sa séquence n'a pas pu être ajoutée à celle de l'élément Alu. Le profil de la bande 1 montre, au niveau du point de cassure, à la fois des

lectures « splittées » de type WT (sans insertion, comme sur le profil de la bande 3) et des lectures qui s'alignent sur la référence avec insertion (allèle mutée). Au niveau de l'insertion, le nombre de lecture alignées est moins important que pour la référence. Le profil de la bande 2 présente un alignement à couverture homogène. On n'observe pas de lectures « splittées » qui indiqueraient la présence de fragments sans insertion. Le profil de la bande 3 présente un défaut d'alignement : aucune lecture ne s'aligne au niveau de l'insertion. Seules des lectures « splittées » sont présentes. Il s'agit de l'allèle WT non porteur de l'insertion Alu. Les profils du père (non présentés) sont identiques à ceux du cas index. Les données issues du séquençage de la mère se comportent comme celles des fragments sans insertion (allèle WT). Les résultats du MiSeq pour le gène *NPRL3* confirment la présence d'une insertion hétérozygote d'un élément identifié comme AluY chez le cas index et son père. La ségrégation familiale et l'existence d'un gène candidat après la lecture recherche conduisent à considérer cette insertion comme de signification inconnue.

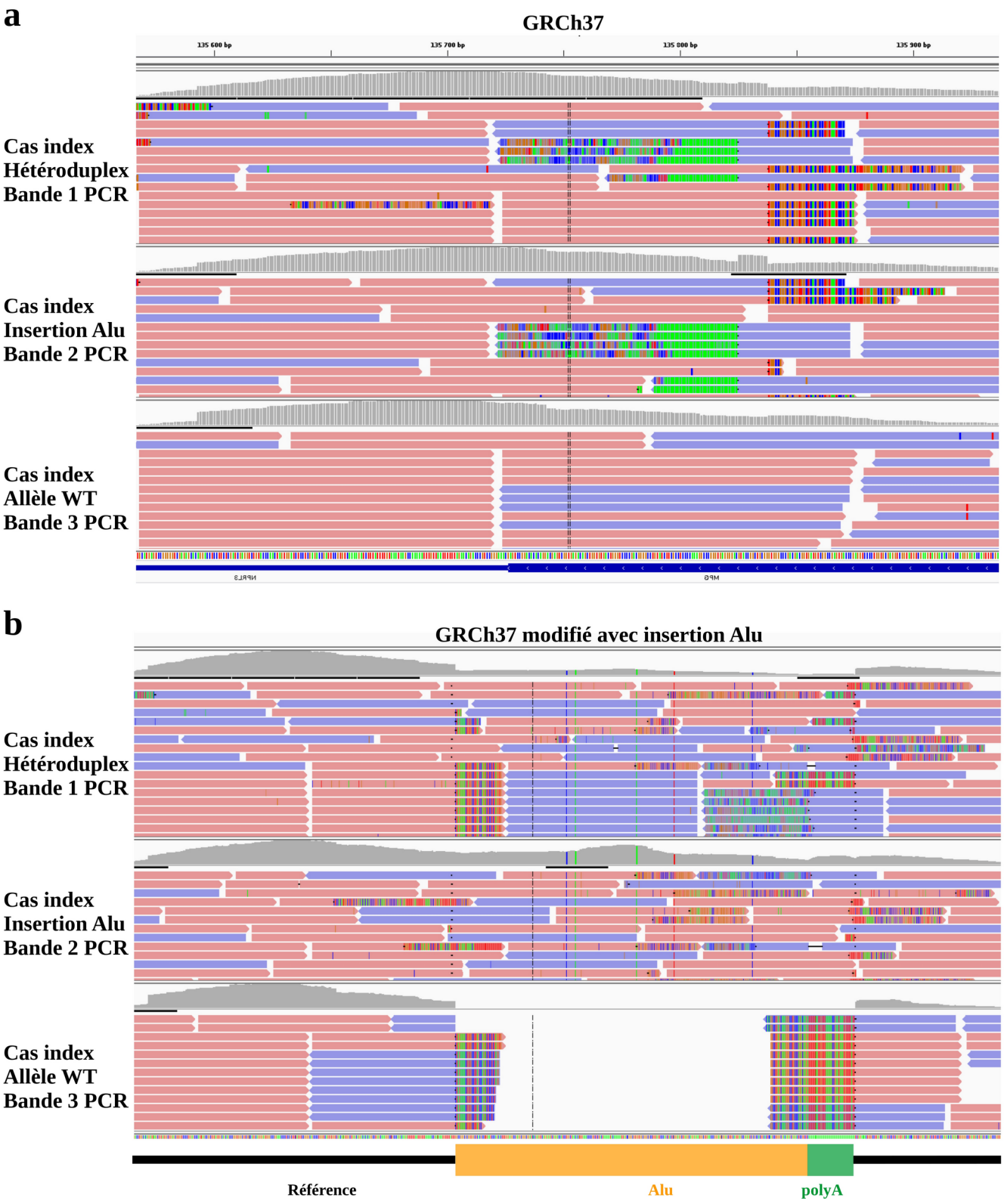

a
GRCh37
135 600 bp
135 700 bp
135 800 bp
135 900 bp
Cas index
Hétéroduplex
Bande 1 PCR
Cas index
Insertion Alu
Bande 2 PCR
Cas index
Allèle WT
Bande 3 PCR
b
GRCh37 modifié avec insertion Alu
Cas index
Hétéroduplex
Bande 1 PCR
Cas index
Insertion Alu
Bande 2 PCR
Cas index
Allèle WT
Bande 3 PCR
Référence
Alu
polyA

Figure 52 : Image IGV de l'insertion de l'élément Alu au sein de la région 3'-UTR du gène NPRL3 chez le cas index

Les 3 produits de PCR obtenus pour le cas index ont été séquencés indépendamment. Les profils sont présentés selon l'ordre de migration (bandes 1, 2 et 3 de la PCR). a) Alignement sur GRCh37. L'insertion d'une région au sein du génome de référence GRCh37 n'apparaît pas sur les profils car elle est absente de la référence. Les bases non-appariées sont représentées en couleur. La région du point de cassure est mise en évidence par les lectures « splittées » qui ne s'alignent que partiellement sur le génome de référence. Profil de la bande 1 : la coexistence de lectures « splittées » et de lectures normales renforce l'hypothèse de l'existence d'hétéroduplex allèle muté/allèle WT. Profil de la bande 2 : l'insertion d'un élément mobile est identifiée par les lectures « splittées ». Profil de la bande 3 : aucune anomalie d'alignement n'est identifiée pour la séquence de l'allèle WT. Les profils du père sont identiques (non présentés). Le profil de la mère est identique (non présenté) à celui de l'allèle WT. b) Alignement sur NPRL3 avec insertion. Profil de la bande 1 : L'alignement des données montre la présence de l'allèle WT et de l'allèle contenant l'insertion Alu et sa queue polyA. Il existe des lectures "splittées" qui ne s'alignent que sur le génome de référence et qui présentent des mésappariements avec la région insérée. Ces lectures sont retrouvées sur le profil de la bande 3. Il existe en parallèle des lectures qui s'alignent sur la référence modifiée malgré la présence de l'insertion. Il y a donc deux fois moins de lectures qui couvrent la région insérée. Profil de la bande 2 : Les données de séquençage s'alignent sur cette référence modifiée par insertion d'un élément Alu. Profil de la bande 3 : L'absence de séquence Alu au sein cet allèle WT est à l'origine de l'absence de lecture alignée au niveau de l'insertion. Seules des lectures « splittées » sont présentes.

IV- DISCUSSION : LA DÉTECTION DES EM, UNE DEUXIÈME LIMITE BIOINFORMATIQUE REPOUSSÉE

Les tests réalisés sur les données 1000 Génomes ont montré que MELT était le meilleur outil pour détection des EM à partir de données de ES. Ces résultats sont cohérents car les données des contrôles ont été obtenus avec la version antérieure de cet outil.

L'analyse de données de séquençage à haut débit d'exome par l'outil MELT a permis de mettre en évidence 3 insertions d'éléments mobiles candidates. Durant cette étude, des défis biologiques et bioinformatiques ont dû être relevés. En effet, les éléments mobiles ne peuvent être identifiés avec le pipeline classique utilisé pour la détection des SNV. Tout d'abord, le module bioinformatique nécessite des données d'alignement (fichiers BAM) et non des données brutes de séquençage (fichiers fastq). De plus, l'identification des éléments mobiles ne passe pas par un appel de variants classique mais par l'étude de paires de lectures particulières.

Une des différences principales est l'utilisation de deux types de séquences de référence. En effet, l'identification des paires de lectures discordantes et des lectures « splittées » nécessite à la fois la référence du génome humain et les séquences des éléments mobiles que l'on souhaite détecter. MELT utilise les séquences consensus des 3 types d'EM et rapporte dans le fichier VCF les différences entre la séquence consensus de l'élément identifié et celle obtenue chez le patient. D'autres outils comme Tangram ou Mobster font appel à une base de données contenant plusieurs sous-familles d'éléments mobiles.

Un autre élément important dans l'identification d'éléments mobiles est la détermination de la position du point de cassure. Sur des données d'exome, seuls ceux présents au sein des régions ciblées peuvent être détectés de manière précise. Par ailleurs cette précision dépend également du nombre de lectures à cette position et surtout de la présence de lectures « splittées » qui restent les meilleurs indicateurs. Néanmoins, malgré la présence de ces dernières, la précision peut varier en fonction des outils utilisés. Ainsi l'élément détecté au sein du gène *GRIN2B* par Mobster et MELT a un point de cassure donné à deux positions différentes : 13 716 609 et 13 716 543 (**Tableau 11**). Mais ces dernières restent proches et inclues dans l'intervalle de confiance à 90 % de [13 716 523-13 716 668]. La caractérisation des EM est donc moins efficace que leur détection mais cette diminution de la précision ne représente pas un obstacle pour la vérification par PCR. De plus, le séquençage MiSeq permet de déterminer avec précision sa localisation. Il serait intéressant d'étudier cette précision à partir des données de génome qui, par définition, couvrent toutes les séquences géniques et permettraient donc d'obtenir une meilleure couverture de ces dernières notamment au niveau des frontières introns-exons. Dans cette optique, l'étude d'une cohorte de 433 individus (cas index + parents) séquencés en génome est en cours d'analyse avec les mêmes pipelines.

Le statut de l'insertion est également déterminé par le programme MELT. Néanmoins, il existe des erreurs d'identification comme le montre le statut hétérozygote détecté pour l'insertion au sein du gène *FERMT1* et dont le statut homozygote a été confirmé par PCR (**Fig. 46**) et MiSeq. La visualisation du fichier BAM initial à cette position avec l'outil IGV montre également le statut homozygote de l'insertion avec uniquement des lectures « splittées » présentes. Le génotype déterminé par l'outil MELT n'est donc pas fiable à 100 %. Il est donc important de tenir compte de cette information lors de l'analyse des résultats ; aussi bien pour

l'étude de la ségrégation en cas de trio que pour la comparaison du génotype avec le mode de transmission des pathologies. L'analyse des éléments mobiles potentiels détectés par MELT a nécessité une annotation des résultats. L'outil MELT fournit les informations concernant l'élément mobile détecté : famille, sous-famille, taille, sens d'insertion, les différences de séquence avec le consensus, la séquence TSD si elle est détectée, le nombre de lectures ayant permis sa détection et les filtres d'analyse qu'il a passés. L'identifiant RefSeq du transcrit correspondant au gène est également indiqué ainsi que la région génique impactée. Néanmoins ces informations ne sont pas suffisantes pour interpréter et filtrer l'ensemble des résultats obtenus. Ainsi, l'annotation par l'outil AnnotSV a été réalisée. Elle permet d'obtenir le nom du gène, phénotype associé sur OMIM s'il existe (numéro OMIM, nom de la pathologie et hérédité) ainsi que les scores bioinformatiques MISZ et pLI du gène. Ces annotations permettent de filtrer l'élément mobile sur le phénotype des patients. D'autres annotations sont également ajoutées et permettent un filtre de fréquence. Il s'agit d'informations provenant des bases de données DDD, DGV, dbVar et 1000 Génomes. Cette dernière répertorie notamment les insertions d'éléments mobiles détectées sur cette population contrôle et en donne la fréquence qui sert de filtre (**Fig. 43**). Les différents filtres appliqués ont ainsi permis d'éliminer les insertions non pertinentes réduisant ainsi de plus de 99 % le nombre d'insertions potentielles à étudier en détail.

Les validations par PCR ont mis en évidence 6 faux-positifs sur 9 candidats (**Figs. 44 et 50**). Ces 6 faux positifs ont été détectés par MELT et ont passé les différents filtres de MELT et de l'interprétation. Mais ils sont absents des fichiers résultats au format BAM produits par MELT et contenant les lectures ayant permis la détection des EM. Pourtant les scores LP, RP et SR, qui indiquent le nombre de lectures en 5', en 3' et « splittées » sur l'élément mobile, ne présentent pas de profils différents de ceux des 3 candidats validés par PCR. Une des questions qui reste en suspens est le lien entre le profil des résultats dans le fichier VCF et leur présence ou absence au sein des fichiers BAM. Des analyses non supervisées : analyses en composantes principales (ACP) ont été réalisées afin de tenter d'identifier les paramètres (nombre de lectures LP, RP ou SR, profondeur, etc.) qui permettraient de distinguer les faux-positifs des vrais-positifs. Aucune des analyses ne s'est révélée concluante. Mais cette étude a été effectuée à

partir d'un faible nombre de résultats. Il serait intéressant de tester un grand nombre de candidats par PCR pour mettre en évidence tous les faux positifs avant de refaire ces ACP. En l'absence de ces résultats, la validation par PCR reste donc une étape importante avant la poursuite de l'analyse individuelle des candidats.

Après validation par PCR de leur présence, 3 éléments mobiles candidats ont été retenus dans les gènes *NPRL3*, *FERMT1* et *GRIN2B*. Les deux premiers ont été séquencés par MiSeq tandis que le séquençage du dernier est en cours. Si la première insertion a été finalement considérée comme de signification inconnue, l'impact des deux autres sur le phénotype des patients nécessitent des investigations supplémentaires notamment des conséquences sur l'ARN afin de confirmer leur caractère pathogène.

Le gène *NPRL3* (OMIM 617118) a été décrit dans l'épilepsie focale familiale à foyers variables, de transmission autosomique dominante. Cette pathologie, à pénétrance incomplète, se manifeste dans la petite enfance. Le patient présente une DI syndromique avec notamment des convulsions tonico-cloniques généralisées. La ségrégation a montré que l'insertion était héritée du père asymptomatique (**Fig. 51**) pour cette pathologie de pénétrance incomplète. Cette insertion a été considérée comme de signification inconnue. L'outil Tangram a également mis en évidence l'insertion chez le cas index mais n'identifie rien chez le père. Selon les résultats de MELT, il existe plus de lectures « splittées » chez le cas index que chez le père. Tangram détectant moins bien les EM, il est possible que le seuil de détection n'ait pas été atteint chez le père. L'outil Mobster ne détecte rien chez les 3 membres du trio. Le séquençage par MiSeq a confirmé qu'un élément Alu s'est inséré au sein de la région 3'-UTR de ce gène. Néanmoins le caractère bénin ou pathogène de cette insertion n'a pas pu être démontré.

Le gène *FERMT1* (OMIM 173650) est impliqué dans le syndrome de Kindler de transmission autosomique récessive. Le patient présente une leucokératose buccale avec poïkilodermie acquise conduisant à l'hypothèse d'une poïkilodermie bulleuse de Kindler. Un élément Alu s'est inséré au sein de l'exon 7 de ce gène. Ni Tangram ni Mobster n'ont mis en évidence cet événement pourtant confirmé par PCR et MiSeq (**Tableau 11**). Le statut homozygote de cette insertion conduit à suspecter une consanguinité chez le patient. Le patient

mentionne que ces parents étaient originaires du même village. La ségrégation familiale montre également le statut homozygote du cas index avec des enfants non atteints hétérozygotes pour cette insertion (**Fig. 46**). Les fragments amplifiés par PCR et contenant l'insertion ont été séquencés par MiSeq pour les 3 enfants. Les fragments sans insertion et les hétéroduplex (bandes 3 et 1 de la migration) ont également été séquencés pour la fille aînée. Les bandes observées sur le profil des 3 enfants correspondent bien aux hétéroduplex, à l'insertion hétérozygote héritée du père et à l'allèle sans insertion héritée de la mère.

L'impact de cette anomalie génétique au niveau transcriptionnel est en cours d'étude sur le tissu cutané affecté.

Le gène *GRIN2B* (OMIM 138252) est impliqué dans l'encéphalopathie épileptique infantile précoce et dans la déficience intellectuelle non syndromique, pathologies de transmission autosomique dominante. La patiente est atteinte de déficience intellectuelle légère, de retard du développement et d'hypotonie. Un élément Alu s'est inséré au sein de l'exon 13 du gène *GRIN2B*. L'outil Mobster détecte également cette insertion chez le cas index tandis que Tangram n'identifie rien à cette position. La ségrégation familiale montre que cette insertion est *de novo* (**Fig. 50**). La forte corrélation entre les informations cliniques et biologiques nous pousse à continuer les investigations avec un séquençage MiSeq et une étude de l'ARN.

Les 3 candidats détectés par MELT n'ont pas tous été identifiés par les 2 autres outils Tangram et Mobster. Tangram a été choisi pour cette analyse car, à l'instar de MELT, il s'agit d'un outil utilisé par le consortium 1000 Génomes pour détecter les éléments transposables de classe I qui utilise à la fois les lectures discordantes (DP) et les lectures « splittées » (SR). Mobster est quant à lui un outil qui repère les éléments Alu, L1, SVA ou HERV-K (Human Endogenous RetroViruses K) en n'utilisant que les DP ce qui permet de comparer l'utilisation des DP couplées aux SR à celle des DP seules. Les résultats obtenus pour les 3 candidats confortent le choix de MELT comme programme d'identification des éléments mobiles pour l'étude de cette cohorte. L'analyse de Gardner et al. a montré en effet que MELT est l'outil de détection des éléments mobiles avec la meilleure sensibilité et la meilleure spécificité (Gardner

et al., 2017). Il serait néanmoins intéressant d'effectuer la même démarche de comparaison en partant des résultats de Tangram puis de Mobster, pour les comparer à ceux des 2 autres outils.

Le premier filtre a été celui des gènes OMIM morbides. Il a permis de diminuer de 84 % le nombre d'éléments mobiles potentiellement candidats. Bien qu'efficace, ce type de filtre a ses limites car il ne permet de se concentrer que sur un nombre restreint de gènes. Une étude des gènes non connus en pathologie humaine pourrait être réalisée sur le modèle de la lecture en recherche des SNV et CNV issus de données d'exome afin d'augmenter les chances d'identifier de nouveaux gènes candidats. Il est également nécessaire d'effectuer périodiquement une mise à jour de l'annotation afin de déceler les éléments mobiles insérés au sein de gènes nouvellement impliqués en pathologie humaine

De plus, pour cette étude, l'analyse des résultats a été concentrée sur les régions non-introniques incluant : exons, 3' et 5'-UTR, promoteurs et terminateurs de transcription. Mais il serait intéressant de pousser l'analyse aux régions introniques grâce au GS. En effet, l'insertion d'un élément mobile entre 2 exons peut avoir un impact sur la transcription avec la création d'un nouveau site d'épissage, l'apparition d'un nouvel exon, l'insertion d'une séquence à transcrire trop importante (LINE-1) pour l'ARN polymérase ou l'apparition d'un signal de polyadénylation prématuré. Ces impacts ne concernent pas uniquement les régions introniques, aussi l'analyse de l'ARN du gène concerné permet de déterminer l'impact de l'insertion de l'élément mobile non seulement sur la séquence du gène en elle-même mais également sur son expression.

Les résultats ont ensuite été filtrés sur la fréquence au sein de la base de données 1000 Génomes (1000 Genomes Project Consortium et al., 2012). Seuls les événements dont la fréquence était inférieure à 1 % ont été conservés afin d'éliminer les polymorphismes (1000 Genomes Project Consortium, 2010). Nous avons également choisi d'appliquer un filtre de fréquence observée au sein de la cohorte. Tout événement d'insertion présent plus de 4 fois au sein des 3322 individus n'a pas été retenu. Une insertion homozygote héritée des 2 parents n'est donc pas éliminée sur ce critère. De même, une éventuelle insertion transmise par les 2 parents à leurs 2 enfants (unique famille de 4 individus de la cohorte) ne sera pas filtrée sur cette fréquence. De plus, l'insertion *de novo* d'un élément mobile à l'origine d'une pathologie représente 0,3 % des variations pathogènes soit 3 pour mille (Cordaux and Batzer, 2009). Des valeurs plus faibles étaient attendues pour des données d'exome qui couvrent une partie plus

restreinte du génome (2 %). L'équipe de Gardner a en effet travaillé sur une cohorte de 9738 trios du projet DDD et a identifié 9 éléments mobiles *de novo* dont 4 probablement pathogènes soit 0,04 % (Gardner et al., 2019) (**Tableau 12**). Une étude publiée au cours de nos travaux retrouve aussi cette valeur avec une cohorte de 38 371 patients atteints majoritairement de retard neuro-développemental (Torene et al., 2020). Treize insertions d'éléments mobiles pathogènes ou probablement pathogènes y ont été identifiées à partir de donnés d'exome et confirmées par Sanger, soit 0,03 %. Notre étude, également réalisée sur des données d'exome, a permis de mettre en évidence 2 candidats dont 1 *de novo* dans une cohorte de 3322 individus dont 2500 cas index. Cette statistique de 1/2500 correspond à celle déterminée lors de ces 2 études précédentes.

Cohorte	Pathologie étudiée	Nombre de cas index inclus	Stratégie utilisée	Résultats *de novo*
Gardner et al., 2019 (cohorte DDD)	Retard du développement	9738	Trio	4 (0,04 %)
Torene et al., 2020	Retard du développement	38371	Trio / solo	13 (0,03%)
GAD	AD + DI	2500	Trio / solo	1 (0,04 %)

Les insertions d'EM identifiées par Gardner et Torene ne sont pas présentes chez les parents dont l'ADN était disponible (confirmées ou présumées *de novo*). Leur fréquence est similaire aux 0,03 % attendus. Dans notre étude, la première insertion (gène *NPRL3*) a été héritée du père et considérée comme de signification inconnue. La seconde (gène *FERMT1*) est supposée héritée des 2 parents et devrait être considérée comme responsable de la pathologie après analyse de l'ARN. La dernière (gène *GRIN2B*) a été confirmée *de novo* et est un candidat solide pour lequel des analyses complémentaires sont en cours. Il y a donc une insertion *de novo* suspectée d'être pathogène pour 2500 cas index au sein de la cohorte (soit 0,04 %).

Tableau 12 : Synthèse des 3 études sur des cohortes de patients atteints d'AD et/ou de DI

Parmi les insertions mises en évidence dans notre cohorte par MELT, 62 étaient des doubles-insertions, c'est-à-dire un élément Alu et un élément L1 détectés au même endroit chez le même patient. La vérification de chaque alignement local a été réalisée par contrôle de la profondeur à l'aide du module DepthOfCoverage de GATK. Il s'est avéré qu'il s'agissait soit de deux faux-positifs (profondeur nulle en raison d'une absence d'alignement) soit d'une insertion et d'un faux-positif ; à l'exception de 2 cas. Ces derniers encadrent une partie du dernier exon

du gène *RPGR*, région de l'exome très peu couverte en raison d'une répétition de C et de T. Il s'agit donc de faux-positifs malgré une profondeur non nulle dans les fichiers BAM générés par MELT.

Outre les éléments L1, Alu et SVA, certaines études considèrent également les éléments HERV, éléments transposables de classe I à LTR. Ces éléments de mobilité très faible (Gifford and Tristem, 2003) se déplaceraient par un mécanisme semblable à ceux des rétrovirus (Turner et al., 2001). Ils sont maintenant reconnus comme cofacteurs de pathologies complexes et sont sensibles aux stimuli externes. De récents travaux, publiés au cours de notre étude, ont mis en évidence un lien entre l'activité anormale de ces éléments lors de l'embryogenèse et l'apparition de pathologies neuro-développementales, sans toutefois prouver une relation de cause à effet (Balestrieri et al., 2019).

Néanmoins, les deux études récentes de Gardner en 2019 puis de Torene en 2020 ne les considèrent pas (Gardner et al., 2019 ; Torene et al., 2020). De plus, MELT est l'outil utilisé par le Consortium 1000 Génomes et semble être le plus efficace. Nous avons choisi de nous concentrer sur les éléments détectés par ce programme : Alu, L1 et SVA.

L'ensemble de ces travaux sur la détection des éléments mobiles sur des données de séquençage d'exome fait l'objet d'une publication en cours de rédaction.

DISCUSSION, CONCLUSION ET PERSPECTIVES

Les travaux réalisés durant cette thèse ont consisté en l'extraction et la réanalyse ou l'optimisation de l'analyse de données préexistantes de ES. Les objectifs ont mené à l'implémentation de pipelines d'analyse et la mise en place de méthodes d'étude de catégories supplémentaires de variations détectables en ES, dans le but de repousser des limites clinico-biologiques et bioinformatiques rencontrées dans l'identification de nouveaux gènes ou de nouveaux mécanismes moléculaires impliqués dans des maladies génétiques rares.

La collaboration avec le Laboratoire Cerba a montré l'importance de l'apport de l'expertise dans la relecture en diagnostic et en recherche de données d'exome. Le partage de ces données (80 exomes négatifs relus par une approche en recherche) a permis l'identification d'un nouveau gène impliqué dans une forme d'encéphalopathie épileptique et la mise en place d'une première collaboration internationale, ainsi que la participation à une autre collaboration internationale pour la caractérisation d'une déficience intellectuelle syndromique. De plus, les nouvelles méthodes mises au point lors de ces travaux pourront être appliquées, par le Laboratoire Cerba, au pipeline d'analyse et à lecture diagnostique des données de ES.

Ces travaux de recherche ont montré qu'il était possible d'extraire des informations supplémentaires à partir des données de ES afin de repousser des limites rencontrées lors de l'analyse d'exome. Chacune des approches décrites, utilisées en routine ou appliquées à la recherche, conduit à l'amélioration du diagnostic et à la diminution de l'errance diagnostique. Ainsi la lecture des exomes dans un cadre de recherche, l'identification de variations mitochondriales et la détection de l'insertion d'éléments mobiles ont permis d'ajouter un total de 5 diagnostics supplémentaires, auxquels pourront s'ajouter les 3 candidats en attente de confirmation (*DLGAP2*, *FERMT1* et *GRIN2B*). Ces méthodes ont été implémentées (analyse de l'ADNmt) ou vont prochainement l'être (détection des EM) au pipeline d'analyse diagnostique, puisque ces travaux ont été réalisés dans un cadre de recherche translationnelle qui vise à optimiser le diagnostic. Il s'agit en effet de transférer systématiquement au domaine clinique les découvertes de la recherche afin qu'elles bénéficient le plus rapidement possible aux patients.

Plusieurs perspectives peuvent être considérées dans l'optique de réduire les situations d'impasse diagnostique, que cela soit dans l'amélioration des connaissances clinico-biologiques, des techniques ou des pipelines bioinformatiques.

Différentes pistes peuvent être évoquées pour optimiser l'analyse des données. Citons la réanalyse des données d'exome, qui grâce à l'augmentation des connaissances, permet d'augmenter de 15 % le taux diagnostique (Nambot et al., 2018 ; Wright et al., 2018). Citons également l'intérêt d'explorer l'utilisation d'un outil de classification des variations, comme Exomiser (Robinson et al., 2014), qui pourrait permettre un gain de temps lors de l'interprétation en seconde intention.

Sur le plan bioinformatique, d'autres perspectives peuvent être citées. Il reste des données non exploitées lors de l'analyse d'exome notamment pour l'ADNmt. En effet, le pipeline d'analyse de l'ADNmt ne permet de détecter que les variations ponctuelles ou les indels de très petites tailles. L'identification des CNV nécessite le développement de méthodes spécifiques supplémentaires comme par exemple le nouvel outil « eKLIPse » (Goudenège et al., 2019). Mais le faible nombre de lectures s'alignant sur les régions capturées indirectement pourrait représenter un frein à ce type d'analyse. En ce qui concerne l'exome nucléaire, il existe des pathologies qui ont pour origine une anomalie génétique autre qu'une variation ponctuelle ou qu'un CNV « classique ». La détection des STR pathogènes constitue en effet une piste supplémentaire pour augmenter le taux diagnostique de l'exome (Tankard et al., 2018).

Sur le plan des progrès moléculaires, l'exome ne représente qu'environ 2 % du génome humain. Il existe donc des variations dans les régions non codantes qui peuvent avoir un impact sur l'expression des gènes, la stabilité de l'ARNm ou la fonction des protéines ; et qui ne sont détectables qu'en génome. D'autres pistes commencent donc à être explorées en combinant bioinformatique, séquençage du génome et analyses multi-omiques. Ainsi, il a été montré que le séquençage à haut débit de génome permet d'augmenter le taux diagnostique (Gilissen et al., 2014). En effet, il donne accès à des régions du génome non disponibles en exome, notamment les régions introniques. Il est maintenant connu que des variations présentes dans ces régions introniques peuvent avoir un effet pathogène et être à l'origine de pathologies (Vaz-Drago et al., 2017). De plus, le séquençage de génome permet de détecter avec précision

des remaniements chromosomiques non identifiables en exome, à l'origine d'une dérégulation de l'expression génique ou d'une interruption de la séquence codante. Les méthodes développées lors de ces travaux de thèse sur les données d'exome sont par ailleurs applicables partiellement aux données de génome. Enfin, le séquençage du génome donne également la possibilité d'identifier par exemple les STR non exoniques, d'augmenter le nombre de lectures pour l'ADNmt ou de mettre en évidence des sites d'épissage cryptiques.

Les analyses multi-omiques se présentent comme de nouvelles possibilités pour résoudre les impasses diagnostiques. Parmi elles, citons l'étude des données d'ARN extraites du sang ou d'un autre tissu pour confirmer l'impact d'une variation sur l'épissage, mettre en évidence un transcrit alternatif, ou mesurer l'expression différentielle d'un ou plusieurs gènes. L'équipe GAD a donc mis en place le projet DIWA qui a pour but de comparer 3 différentes approches : (i) génome en solo + RNA-seq, (ii) génome en trio et (iii) génome en trio + RNA-seq. L'objectif est de déterminer l'apport de ces technologies dans l'identification des causes génétiques des pathologies rares avec AD/DI. Toujours dans la même optique, l'équipe a également le projet OMIXCARE qui vise à comparer le séquençage de fragments courts et longs couplés au RNA-seq.

Les analyses du métabolome sur sang ou urine, de l'épigénome, du protéome ou l'étude de modèles cellulaires de type iPS sont également d'autres possibilités qui peuvent être envisagées (**Fig. 53**). Ainsi, l'équipe GAD est impliquée dans le projet Solve-RD notamment en ce qui concerne l'apport des stratégies multi-omiques dans l'identification de nouveaux mécanismes à l'origine de pathologies génétiques rares. Mais l'intégration de l'ensemble de ces informations aux données de séquençage reste encore un défi.

Néanmoins, les améliorations constantes des techniques et des analyses de données de séquençage à haut débit font de l'exome et du génome des outils puissants pour la recherche de nouveaux mécanismes moléculaires impliqués dans des maladies génétiques rares.

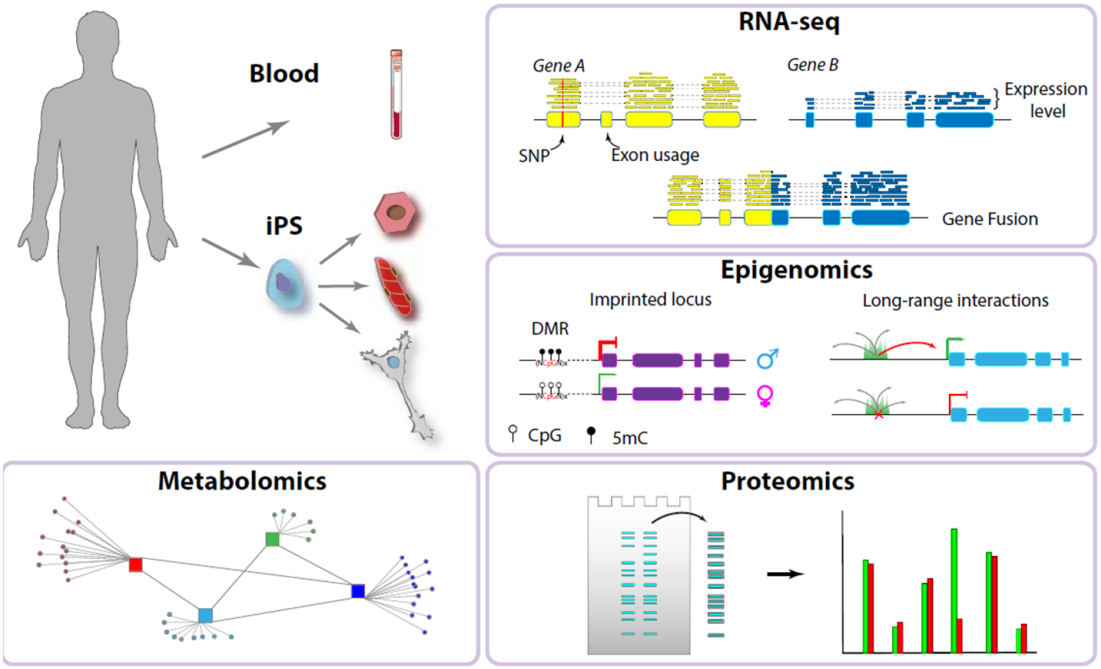


Figure 53 : Approches multi-omiques dans l'identification de nouveaux mécanismes à l'origine de pathologies génétiques rares
Les iPS permettent de réaliser des études sur des lignées cellulaires très diverses à partir de prélèvements plus aisés à obtenir. Le RNA-seq donne accès au transcriptome et aux modifications de l'expression génique causées par une variation. L'épigénome permet de détecter les changements de marques épigénétiques et de déterminer l'impact d'une variation sur la régulation génique. Le métabolome renseigne sur l'impact d'une variation sur l'activité d'une cellule. Chaque type de cellules possède en effet sa propre « empreinte ». Le protéome, en lien avec le métabolome, renseigne sur le fonctionnement cellulaire et sur les modifications pré- et post-traductionnelles (figure crée par Antonio Vitobello).

RÉFÉRENCES

1000 Genomes Project (2015). A global reference for human genetic variation. Nature *526*, 68.

1000 Genomes Project Consortium (2010). A map of human genome variation from population scale sequencing. Nature *467*, 1061–1073.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Adav, S.S., and Sze, S.K. (2016). Insight of brain degenerative protein modifications in the pathology of neurodegeneration and dementia by proteomic profiling. Mol Brain *9*.

Aicardi, J. (1998). The etiology of developmental delay. Seminars in Pediatric Neurology *5*, 15–20.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell (Garland Science).

Al-Nabhani, M., Al-Rashdi, S., Al-Murshedi, F., Al-Kindi, A., Al-Thihli, K., Al-Saegh, A., Al-Futaisi, A., Al-Mamari, W., Zadjali, F., and Al-Maawali, A. (2018). Reanalysis of exome sequencing data of intellectual disability samples: Yields and benefits. Clinical Genetics *94*, 495–501.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available Online at: Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics *23*, 147.

Badano, J.L., and Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet *3*, 779–789.

Bai, R.-K., and Wong, L.-J.C. (2004). Detection and Quantification of Heteroplasmic Mutant Mitochondrial DNA by Real-Time Amplification Refractory Mutation System Quantitative PCR Analysis: A Single-Step Approach. Clinical Chemistry *50*, 996–1001.

Bai, R.-K., and Wong, L.-J.C. (2005). Simultaneous Detection and Quantification of Mitochondrial DNA Deletion(s), Depletion, and Over-Replication in Patients with Mitochondrial Disease. The Journal of Molecular Diagnostics *7*, 613–622.

Bailey, J.A., Carrel, L., Chakravarti, A., and Eichler, E.E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis. Proc Natl Acad Sci U S A *97*, 6634–6639.

Balestrieri, E., Matteucci, C., Cipriani, C., Grelli, S., Ricceri, L., Calamandrei, G., and Sinibaldi Vallebona, P. (2019). Endogenous Retroviruses Activity as a Molecular Signature of Neurodevelopmental Disorders. Int J Mol Sci *20*.

Balogh, K., Patócs, A., Majnik, J., Rácz, K., and Hunyady, L. (2004). Genetic screening methods for the detection of mutations responsible for multiple endocrine neoplasia type 1. Molecular Genetics and Metabolism *83*, 74–81.

Bannwarth, S., Procaccio, V., and Paquis-Flucklinger, V. (2005). Surveyor Nuclease: a new strategy for a rapid identification of heteroplasmic mitochondrial DNA mutations in patients with respiratory chain defects. Hum. Mutat. *25*, 575–582.

Bannwarth, S., Procaccio, V., Lebre, A.S., Jardel, C., Chaussenot, A., Hoarau, C., Maoulida, H., Charrier, N., Gai, X., Xie, H.M., et al. (2013). Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders. Journal of Medical Genetics jmedgenet-2013-101604.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA *6*.

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 Elements in Structural Variation and Disease. Annu Rev Genomics Hum Genet *12*, 187–215.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. Nature *456*, 53–59.

Bergant, G., Maver, A., Lovrecic, L., Čuturilo, G., Hodzic, A., and Peterlin, B. (2018). Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. Genet Med *20*, 303–312.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics *28*, 423–425.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bonneau, D., Colin, E., Oca, F., Ferré, M., Chevrollier, A., Guéguen, N., Desquiret-Dumas, V., N'Guyen, S., Barth, M., Zanlonghi, X., et al. (2014). Early-onset Behr syndrome due to compound heterozygous mutations in OPA1. Brain *137*, e301–e301.

Boucret, L., Bris, C., Seegers, V., Goudenège, D., Desquiret-Dumas, V., Domin-Bernhard, M., Ferré-L'Hotellier, V., Bouet, P.E., Descamps, P., Reynier, P., et al. (2017). Deep sequencing shows that oocytes are not prone to accumulate mtDNA heteroplasmic mutations during ovarian ageing. Hum Reprod *32*, 2101–2109.

Broad Institute. (n.d.). Picard Tools. Retrieved from http://broadinstitute.github.io/picard/

Bruel, A.-L., Vitobello, A., Tran Mau-Them, F., Nambot, S., Sorlin, A., Denommé-Pichon, A.-S., Delanne, J., Moutton, S., Callier, P., Duffourd, Y., et al. (2020). Next-Generation Sequencing

approaches and challenges in the diagnosis of developmental abnormalities and intellectual disability. Clinical Genetics.

Bruel, A.-L., Nambot, S., Quéré, V., Vitobello, A., Thevenon, J., Assoum, M., Moutton, S., Houcinat, N., Lehalle, D., Jean-Marçais, N., et al. (2019a). Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. Eur J Hum Genet 1–13.

Bruel, A.-L., Vitobello, A., Mau-Them, F.T., Nambot, S., Duffourd, Y., Quéré, V., Kuentz, P., Garret, P., Thevenon, J., Moutton, S., et al. (2019b). 2.5 years' experience of GeneMatcher data-sharing: a powerful tool for identifying new genes responsible for rare diseases. Genet Med *21*, 1657–1661.

Buermans, H.P.J., and den Dunnen, J.T. (2014). Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease *1842*, 1932–1941.

Bumgarner, R. (2013). DNA microarrays: Types, Applications and their future. Curr Protoc Mol Biol *0 22*, Unit-22.1.

Calabrese, F.M., Simone, D., and Attimonelli, M. (2012). Primates and mouse NumtS in the UCSC Genome Browser. BMC Bioinformatics *13*, S15.

Calabrese, F.M., Clima, R., Pignataro, P., Lasorsa, V.A., Hogarty, M.D., Castellano, A., Conte, M., Tonini, G.P., Iolascon, A., Gasparre, G., et al. (2016). A comprehensive characterization of rare mitochondrial DNA variants in neuroblastoma. Oncotarget *7*, 49246–49258.

Chen, J.-M., Cooper, D.N., Chuzhanova, N., Férec, C., and Patrinos, G.P. (2007). Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet *8*, 762–775.

Chen, L.-L., DeCerbo, J.N., and Carmichael, G.G. (2008). Alu element-mediated gene silencing. EMBO J *27*, 1694–1705.

Chenais, B. (2015). Transposable Elements in Cancer and Other Human Diseases. Current Cancer Drug Targets *15*, 227–242.

Chien, W.-H., Gau, S.-F., Wu, Y.-Y., Huang, Y.-S., Fang, J.-S., Chen, Y.-J., Soong, W.-T., Chiu, Y.-N., and Chen, C.-H. (2010). Identification and molecular characterization of two novel chromosomal deletions associated with autism. Clinical Genetics *78*, 449–456.

Chien, W.-H., Gau, S.S.-F., Liao, H.-M., Chiu, Y.-N., Wu, Y.-Y., Huang, Y.-S., Tsai, W.-C., Tsai, H.-M., and Chen, C.-H. (2013). Deep exon resequencing of DLGAP2 as a candidate gene of autism spectrum disorders. Mol Autism *4*, 26.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A *106*, 19096–19101.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) *6*, 80–92.

Clark, M.J., Chen, R., Lam, H.Y.K., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J., and Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. Nat Biotechnol *29*, 908–914.

Collins, F.S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. Science *300*, 286–290.

Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Human Genetics *132*, 1077.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet *10*, 691–703.

Cui, H., Li, F., Chen, D., Wang, G., Truong, C.K., Enns, G.M., Graham, B., Milone, M., Landsverk, M.L., Wang, J., et al. (2013). Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. Genet. Med. *15*, 388–394.

Damert, A., Raiz, J., Horn, A.V., Löwer, J., Wang, H., Xing, J., Batzer, M.A., Löwer, R., and Schumann, G.G. (2009). 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res *19*, 1992–2008.

Deininger, P.L., and Batzer, M.A. (1999). Alu Repeats and Human Disease. Molecular Genetics and Metabolism *67*, 183–193.

Dinwiddie, D.L., Smith, L.D., Miller, N.A., Atherton, A.M., Farrow, E.G., Strenk, M.E., Soden, S.E., Saunders, C.J., and Kingsmore, S.F. (2013). Diagnosis of mitochondrial disorders by concomitant next-generation sequencing of the exome and mitochondrial genome. Genomics *102*, 148–156.

Diroma, M.A., Calabrese, C., Simone, D., Santorsola, M., Calabrese, F.M., Gasparre, G., and Attimonelli, M. (2014). Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. BMC Genomics *15*, S2.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., et al. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics *35*, 4754–4756.

Dracopoulos, A., Widjaja, E., Raybaud, C., Westall, C.A., and Snead, O.C. (2010). Vigabatrin-associated reversible MRI signal changes in patients with infantile spasms. Epilepsia *51*, 1297–1304.

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*.

Endris, V., Hackmann, K., Neuhann, T.M., Grasshoff, U., Bonin, M., Haug, U., Hahn, G., Schallner, J.C., Schröck, E., Tinschert, S., et al. (2010). Homozygous loss of CHRNA7 on chromosome 15q13.3 causes severe encephalopathy with seizures and hypotonia. American Journal of Medical Genetics Part A *152A*, 2908–2911.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat Genet *24*, 363–367.

Fan, H., and Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat Mutation. Genomics Proteomics Bioinformatics *5*, 7–14.

Farwell, K.D., Shahmirzadi, L., El-Khechen, D., Powis, Z., Chao, E.C., Tippin Davis, B., Baxter, R.M., Zeng, W., Mroske, C., Parra, M.C., et al. (2015). Enhanced utility of family-centered diagnostic exome sequencing with inheritance model–based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genetics in Medicine *17*, 578–586.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. Nat Genet *41*, 563–571.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nature Reviews Genetics *7*, 85–97.

Finnegan, D.J. (2012). Retrotransposons. Current Biology *22*, R432–R437.

Fondation Maladies Rares La définition des maladies rares | Fondation maladies rares.

Friedman, J.M., Baross, Á., Delaney, A.D., Ally, A., Arbour, L., Asano, J., Bailey, D.K., Barber, S., Birch, P., Brown-John, M., et al. (2006). Oligonucleotide Microarray Analysis of Genomic Imbalance in Children with Mental Retardation. Am J Hum Genet *79*, 500–513.

Fromer, M., and Purcell, S.M. (2014). Using XHMM software to detect copy number variation in whole-exome sequencing data. Curr Protoc Hum Genet *81*, 7.23.1-7.23.21.

Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res *27*, 1916–1929.

Gardner, E.J., Prigmore, E., Gallone, G., Danecek, P., Samocha, K.E., Handsaker, J., Gerety, S.S., Ironfield, H., Short, P.J., Sifrim, A., et al. (2019). Contribution of retrotransposition to developmental disorders. Nat Commun *10*, 1–10.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. J Mol Biol *357*, 1383–1393.

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics *34*, 3572–3574.

Ghosh, A.K., Majumder, M., Steele, R., White, R.A., and Ray, R.B. (2001). A Novel 16-Kilodalton Cellular Protein Physically Interacts with and Antagonizes the Functional Activity of c-myc Promoter-Binding Protein 1. Mol Cell Biol *21*, 655–662.

Gifford, R., and Tristem, M. (2003). The Evolution, Distribution and Diversity of Endogenous Retroviruses. 26.

Giles, R.E., Blanc, H., Cann, H.M., and Wallace, D.C. (1980). Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A *77*, 6715–6719.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature *511*, 344–347.

Goerner-Potvin, P., and Bourque, G. (2018). Computational tools to unmask transposable elements. Nature Reviews Genetics *19*, 688–704.

Golan, D., and Medvedev, P. (2013). Using state machines to model the Ion Torrent sequencing process and to improve read error rates. Bioinformatics *29*, i344–i351.

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

Gotea, V., and Makalowski, W. (2006). Do transposable elements really contribute to proteomes? Trends in Genetics *22*, 260–267.

Goudenège, D., Bris, C., Hoffmann, V., Desquiret-Dumas, V., Jardel, C., Rucheton, B., Bannwarth, S., Paquis-Flucklinger, V., Lebre, A.S., Colin, E., et al. (2019). eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. Genet Med *21*, 1407–1416.

Greally, J.M. (2002). Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc Natl Acad Sci U S A *99*, 327–332.

Gregory, T.R. (2005). Synergy between sequence and size in Large-scale genomics. Nat Rev Genet *6*, 699–708.

Griffin, H.R., Pyle, A., Blakely, E.L., Alston, C.L., Duff, J., Hudson, G., Horvath, R., Wilson, I.J., Santibanez-Koref, M., Taylor, R.W., et al. (2014). Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. Genet Med *16*, 962–971.

Guilmatre, A., Huguet, G., Delorme, R., and Bourgeron, T. (2014). The emerging role of SHANK genes in neuropsychiatric disorders: SHANK Genes in Neuropsychiatric Disorders. Developmental Neurobiology *74*, 113–122.

Han, J.S., Szak, S.T., and Boeke, J.D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature *429*, 268–274.

Hancks, D.C., and Kazazian, H. (2010). SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol *20*, 234–245.

Hancks, D.C., and Kazazian, H.H. (2016). Roles for retrotransposon insertions in human disease. Mob DNA *7*.

Hayakawa, T., Satta, Y., Gagneux, P., Varki, A., and Takahata, N. (2001). Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. PNAS *98*, 11399–11404.

Holt, I.J., Harding, A.E., Petty, R.K., and Morgan-Hughes, J.A. (1990). A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. Am J Hum Genet *46*, 428–433.

Homan, C.C., Kumar, R., Nguyen, L.S., Haan, E., Raymond, F.L., Abidi, F., Raynaud, M., Schwartz, C.E., Wood, S.A., Gecz, J., et al. (2014). Mutations in USP9X are associated with X-linked intellectual disability and disrupt neuronal cell migration and growth. Am. J. Hum. Genet. *94*, 470–478.

Hu, H., Haas, S.A., Chelly, J., Van Esch, H., Raynaud, M., de Brouwer, A.P.M., Weinert, S., Froyen, G., Frints, S.G.M., Laumonnier, F., et al. (2016). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. Mol Psychiatry *21*, 133–148.

Hu, M., Jex, A.R., Campbell, B.E., and Gasser, R.B. (2007). Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. Nature Protocols *2*, 2339–2344.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Jiang-Xie, L.-F., Liao, H.-M., Chen, C.-H., Chen, Y.-T., Ho, S.-Y., Lu, D.-H., Lee, L.-J., Liou, H.-H., Fu, W.-M., and Gau, S.S.-F. (2014). Autism-associated gene Dlgap2 mutant mice demonstrate exacerbated aggressive behaviors and orbitofrontal cortex deficits. Molecular Autism *5*, 32.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol *30*, 772–780.

Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988). Haemophilia A resulting from de novo insertion of L 1 sequences represents a novel mechanism for mutation in man. Nature *332*, 164–166.

Kim, E., Naisbitt, S., Hsueh, Y.-P., Rao, A., Rothschild, A., Craig, A.M., and Sheng, M. (1997). GKAP, a Novel Synaptic Protein That Interacts with the Guanylate Kinase-like Domain of the PSD-95/SAP90 Family of Channel Clustering Molecules. The Journal of Cell Biology *136*, 669–678.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. (1998). Transposable Elements and Genome Organization: A Comprehensive Survey of Retrotransposons Revealed by the Complete Saccharomyces cerevisiae Genome Sequence. Genome Res. *8*, 464–478.

Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A.W., Kleefstra, T., Kramer, J.M., et al. (2016). Systematic Phenomics Analysis

Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. The American Journal of Human Genetics *98*, 149–164.

Komander, D., Clague, M.J., and Urbé, S. (2009). Breaking the chains: structure and function of the deubiquitinases. Nature Reviews Molecular Cell Biology *10*, 550–563.

Kondo-Iida, E., Kobayashi, K., Watanabe, M., Sasaki, J., Kumagai, T., Koide, H., Saito, K., Osawa, M., Nakamura, Y., and Toda, T. (1999). Novel Mutations and Genotype-Phenotype Relationships in 107 Families With Fukuyama-Type Congenital Muscular Dystrophy (FCMD). Hum Mol Genet *8*, 2303–2309.

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C.E., Albarca Aguilera, M., Meyer, R., and Massouras, A. (2019). VarSome: the human genomic variant search engine. Bioinformatics *35*, 1978–1980.

Kowalski, J.R., and Juo, P. (2012). The Role of Deubiquitinating Enzymes in Synaptic Function and Nervous System Diseases. Neural Plast *2012*.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062–D1067.

Laver, T.W., Franco, E.D., Johnson, M.B., Patel, K., Ellard, S., Weedon, M.N., Flanagan, S.E., and Wakeling, M.N. (2019). SavvyCNV: genome-wide CNV calling from off-target reads. BioRxiv 617605.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biology *15*, R84.

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. (2014a). Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. JAMA *312*, 1880–1887.

Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., and Marth, G.T. (2014b). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLoS One *9*.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lelieveld, S.H., Veltman, J.A., and Gilissen, C. (2016). Novel bioinformatic developments for exome sequencing. Hum Genet *135*, 603–614.

Leong, I., Skinner, J., and Love, D. (2014). Application of Massively Parallel Sequencing in the Clinical Diagnostic Testing of Inherited Cardiac Conditions. Medical Sciences *2*, 98–126.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J.-M., Lu, C.-L., Cheng, M.-C., Luu, S.-U., Hsu, S.-H., Hu, T.-M., Tsai, H.-Y., and Chen, C.-H. (2014). Role of the DLGAP2 Gene Encoding the SAP90/PSD-95-Associated Protein 2 in Schizophrenia. PLoS One *9*.

de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N Engl J Med *367*, 1921–1929.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems.

Liu, Z., Zhu, L., Roberts, R., and Tong, W. (2019). Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We? Trends in Genetics *35*, 852–867.

Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R., and Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol *10*, 599–606.

Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D.C. (2013). mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. Curr Protoc Bioinformatics *1*, 1.23.1-1.23.26.

Loublier, S., Schiff, M., Bénit, P., and Rustin, P. (2009). Les maladies mitochondriales : une médecine à part ? Immuno-Analyse & Biologie Spécialisée *24*, 240–253.

Lowther, C., Costain, G., Stavropoulos, D.J., Melvin, R., Silversides, C.K., Andrade, D.M., So, J., Faghfoury, H., Lionel, A.C., Marshall, C.R., et al. (2015). Delineating the 15q13.3 microdeletion phenotype: a case series and comprehensive review of the literature. Genet Med *17*, 149–157.

Luo, S., Valencia, C.A., Zhang, J., Lee, N.-C., Slone, J., Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., et al. (2018). Biparental Inheritance of Mitochondrial DNA in Humans. Proc Natl Acad Sci U S A *115*, 13039–13044.

Ma, Z., Lee, R.W., Li, B., Kenney, P., Wang, Y., Erikson, J., Goyal, S., and Lao, K. (2013). Isothermal amplification method for next-generation sequencing. Proceedings of the National Academy of Sciences *110*, 14320–14323.

MacKinnon, R.N., Selan, C., Zordan, A., Wall, M., Nandurkar, H., and Campbell, L.J. (2012). CGH and SNP array using DNA extracted from fixed cytogenetic preparations and long-term refrigerated bone marrow specimens. Mol Cytogenet *5*, 10.

Maitra, A., Cohen, Y., Gillespie, S.E.D., Mambo, E., Fukushima, N., Hoque, M.O., Shah, N., Goggins, M., Califano, J., Sidransky, D., et al. (2004). The Human MitoChip: A High-Throughput Sequencing Microarray for Mitochondrial Mutation Detection. Genome Res *14*, 812–819.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., and Jabado, N. (2011). What can exome sequencing do for you? Journal of Medical Genetics *48*, 580–589.

Maneechay, W., Boonpipattanapong, T., Kanngurn, S., Puttawibul, P., Geater, S.L., and Sangkhathat, S. (2015). Single Nucleotide Polymorphisms in the Gc Gene for Vitamin D Binding Protein in Common Cancers in Thailand. Asian Pacific Journal of Cancer Prevention *16*, 3339–3344.

Manheimer, K.B., Richter, F., Edelmann, L.J., D'Souza, S.L., Shi, L., Shen, Y., Homsy, J., Boskovski, M.T., Tai, A.C., Gorham, J., et al. (2018). Robust identification of mosaic variants in congenital heart disease. Hum Genet *137*, 183–193.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. Trends in Genetics *24*, 133–141.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature *437*, 376–380.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., et al. (2008). Structural Variation of Chromosomes in Autism Spectrum Disorder. Am J Hum Genet *82*, 477–488.

Masurel-Paulet, A., Andrieux, J., Callier, P., Cuisset, J., Le Caignec, C., Holder, M., Thauvin-Robinet, C., Doray, B., Flori, E., Alex-Cordier, M., et al. (2010). Delineation of 15q13.3 microdeletions. Clinical Genetics *78*, 149–161.

Masurel-Paulet, A., Drumare, I., Holder, M., Cuisset, J.-M., Vallée, L., Defoort, S., Bourgois, B., Pernes, P., Cuvellier, J.-C., Huet, F., et al. (2014). Further delineation of eye manifestations in homozygous 15q13.3 microdeletions including TRPM1: A differential diagnosis of ceroid lipofuscinosis. American Journal of Medical Genetics Part A *164*, 1537–1544.

Matsumoto, A., Mizuno, M., Hamada, N., Nozaki, Y., Jimbo, E.F., Momoi, M.Y., Nagata, K., and Yamagata, T. (2014). LIN7A Depletion Disrupts Cerebral Cortex Development, Contributing to Intellectual Disability in 12q21-Deletion Syndrome. PLoS One *9*.

Mazzarotto, F., Olivotto, I., and Walsh, R. (2020). Advantages and Perils of Clinical Whole-Exome and Whole-Genome Sequencing in Cardiomyopathy. Cardiovasc Drugs Ther.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297–1303.

Meischl, C., de Boer, M., Åhlin, A., and Roos, D. (2000). A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. Eur J Hum Genet *8*, 697–703.

Mevissen, T.E.T., Hospenthal, M.K., Geurink, P.P., Elliott, P.R., Akutsu, M., Arnaudo, N., Ekkebus, R., Kulathu, Y., Wauer, T., El Oualid, F., et al. (2013). OTU Deubiquitinases Reveal

Mechanisms of Linkage Specificity and Enable Ubiquitin Chain Restriction Analysis. Cell *154*, 169–184.

Mevissen, T.E.T., Kulathu, Y., Mulder, M.P.C., Geurink, P.P., Maslen, S.L., Gersch, M., Elliott, P.R., Burke, J.E., van Tol, B.D.M., Akutsu, M., et al. (2016). Molecular basis of Lys11-polyubiquitin specificity in the deubiquitinase Cezanne. Nature *538*, 402–405.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res *16*, 1182–1190.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet *31*, 159–165.

Mullaney, J.M., Mills, R.E., Pittard, W.S., and Devine, S.E. (2010). Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet *19*, R131–R136.

Naisbitt, S., Valtschanoff, J., Allison, D.W., Sala, C., Kim, E., Craig, A.M., Weinberg, R.J., and Sheng, M. (2000). Interaction of the Postsynaptic Density-95/Guanylate Kinase Domain-Associated Protein Complex with a Light Chain of Myosin-V and Dynein. J. Neurosci. *20*, 4524–4534.

Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature *478*, 57–63.

Nambot, S., Thevenon, J., Kuentz, P., Duffourd, Y., Tisserant, E., Bruel, A.-L., Mosca-Boidron, A.-L., Masurel-Paulet, A., Lehalle, D., Jean-Marçais, N., et al. (2018). Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. Genet Med *20*, 645–654.

Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K.V., McDonald, M.T., Meisler, M.H., and Goldstein, D.B. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. J Med Genet *49*, 353–361.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. Nature *461*, 272–276.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet *42*, 30–35.

NIH U.S. National Library of Medecine. (n.d.). Retrieved from https://ghr.nlm.nih.gov/primer/inheritance/inheritancepatterns

Nozu, K., Iijima, K., Ohtsuka, Y., Fu, X.J., Kaito, H., Nakanishi, K., and Vorechovsky, I. (2014). Alport syndrome caused by a COL4A5 deletion and exonization of an adjacent AluY. Mol Genet Genomic Med *2*, 451–453.

Nyren, P., Pettersson, B., and Uhlen, M. (1993). Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. Analytical Biochemistry *208*, 171–175.

Okamoto, K., and Shaw, J.M. (2005). Mitochondrial Morphology and Dynamics in Yeast and Multicellular Eukaryotes. Annual Review of Genetics *39*, 503–536.

OMIM - Online Mendelian Inheritance in Man. (1996). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).

Ostertag, E.M., and Kazazian Jr, H.H. (2001). Biology of Mammalian L1 Retrotransposons. Annual Review of Genetics *35*, 501–538.

Ozgen, H.M., Daalen, E.V., Bolton, P.F., Maloney, V.K., Huang, S., Cresswell, L., Boogaard, M.V.D., Eleveld, M.J., Slot, R.V., Hochstenbach, R., et al. (2009). Copy number changes of the microcephalin 1 gene (MCPH1) in patients with autism spectrum disorders. Clinical Genetics *76*, 348–356.

Patowary, A., Nesbitt, R., Archer, M., Bernier, R., and Brkanac, Z. (2017). Next Generation Sequencing Mitochondrial DNA Analysis in Autism Spectrum Disorder. Autism Res *10*, 1338–1343.

Pfeffer, G., Blakely, E.L., Alston, C.L., Hassani, A., Boggild, M., Horvath, R., Samuels, D.C., Taylor, R.W., and Chinnery, P.F. (2012). Adult-onset spinocerebellar ataxia syndromes due to MTATP6 mutations. J Neurol Neurosurg Psychiatry *83*, 883–886.

Picardi, E., and Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. Nature Methods *9*, 523–524.

Pickering, A.M., Koop, A.L., Teoh, C.Y., Ermak, G., Grune, T., and Davies, K.J.A. (2010). The Immunoproteasome, The 20S Proteasome, And The PA28αβ Proteasome Regulator Are Oxidative-Stress-Adaptive Proteolytic Complexes. Biochem J *432*, 585–594.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorder. Nature *466*, 368–372.

Pinto-Fernández, A., Davis, S., Schofield, A.B., Scott, H.C., Zhang, P., Salah, E., Mathea, S., Charles, P.D., Damianou, A., Bond, G., et al. (2019). Comprehensive Landscape of Active Deubiquitinating Enzymes Profiled by Advanced Chemoproteomics. Front Chem *7*, 592.

Polak, P., and Domany, E. (2006). Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics *7*, 133.

Poquet, H., Faivre, L., Chehadeh, S.E., Morton, J., McMullan, D., Hamilton, S., Goel, H., Isidor, B., Caignec, C.L., Andrieux, J., et al. (2017). Further Evidence for Dlgap2 as Strong Autism Spectrum Disorders/Intellectual Disability Candidate Gene. Autism-Open Access *07*.

Qi, C., Liu, S., Qin, R., Zhang, Y., Wang, G., Shang, Y., Wang, Y., and Liang, J. (2014). Coordinated Regulation of Dendrite Arborization by Epigenetic Factors CDYL and EZH2. J Neurosci *34*, 4494–4508.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Löwer, J., Strätling, W.H., Löwer, R., and Schumann, G.G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res *40*, 1666–1683.

Rasmussen, A.H., Rasmussen, H.B., and Silahtaroglu, A. (2017). The DLGAP family: neuronal expression, function and role in brain disorders. Molecular Brain *10*.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. The Lancet *380*, 1674–1682.

Reeg, S., and Grune, T. (2015). Protein Oxidation in Aging: Does It Play a Role in Aging Progression? Antioxid Redox Signal *23*, 239–255.

Rius, R., Cowley, M.J., Riley, L., Puttick, C., Thorburn, D.R., and Christodoulou, J. (2019). Biparental inheritance of mitochondrial DNA in humans is not a common phenomenon. Genet Med *21*, 2823–2826.

Robinson, P.N., Köhler, S., Oellrich, A., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res *24*, 340–348.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A Sequencing Method Based on Real-Time Pyrophosphate. Science *281*, 363–365.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Rousseau, A., and Bertolotti, A. (2018). Regulation of proteasome assembly and activity in health and disease. Nat Rev Mol Cell Biol *19*, 697–712.

Samuels, D.C., Han, L., Li, J., Quanghu, S., Clark, T.A., Shyr, Y., and Guo, Y. (2013). Finding the lost treasures in exome sequencing data. Trends Genet *29*, 593–599.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. *94*, 441–448.

Santiago-Sim, T., Burrage, L.C., Ebstein, F., Tokita, M.J., Miller, M., Bi, W., Braxton, A.A., Rosenfeld, J.A., Shahrour, M., Lehmann, A., et al. (2017). Biallelic Variants in OTUD6B Cause an Intellectual Disability Syndrome Associated with Seizures and Dysmorphic Features. Am. J. Hum. Genet. *100*, 676–688.

Santorsola, M., Calabrese, C., Girolimetti, G., Diroma, M.A., Gasparre, G., and Attimonelli, M. (2016). A multi-parametric workflow for the prioritization of mitochondrial DNA variants of clinical interest. Hum Genet *135*, 121–136.

Sato, M., and Sato, K. (2013). Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *1833*, 1979–1984.

Schob, C., Morellini, F., Ohana, O., Bakota, L., Hrynchak, M.V., Brandt, R., Brockmann, M.D., Cichon, N., Hartung, H., Hanganu-Opatz, I.L., et al. (2019). Cognitive impairment and autistic-like behaviour in SAPAP4-deficient mice. Transl Psychiatry *9*.

Sen, S.K., Huang, C.T., Han, K., and Batzer, M.A. (2007). Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res *35*, 3741–3751.

Shankar, R., Grover, D., Brahmachari, S.K., and Mukerji, M. (2004). Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. BMC Evol Biol *4*, 37.

Sharp, A.J., Mefford, H.C., Li, K., Baker, C., Skinner, C., Stevenson, R.E., Schroer, R.J., Novara, F., De Gregori, M., Ciccone, R., et al. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet *40*, 322–328.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol *26*, 1135–1145.

Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308–311.

Simon, J., Stoll, K., Fick, R., Mott, J., and Lawson-Yuen, A. (2019). Homozygous 15q13.3 microdeletion in a child with hypotonia and impaired vision: A new report and review of the literature. Clinical Case Reports *n/a*.

Slotkin, W., and Nishikura, K. (2013). Adenosine-to-inosine RNA editing and human disease. Genome Medicine *5*, 105.

Smit, A., Hubley, R., and Green, P. (2008). RepeatMasker Home Page.

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.

Soler-Alfonso, C., Carvalho, C.M., Ge, J., Roney, E.K., Bader, P.I., Kolodziejska, K.E., Miller, R.M., Lupski, J.R., Stankiewicz, P., Cheung, S.W., et al. (2014). CHRNA7 triplication associated with cognitive impairment and neuropsychiatric phenotypes in a three-generation pedigree. Eur J Hum Genet *22*, 1071–1076.

Sonney, S., Leipzig, J., Lott, M.T., Zhang, S., Procaccio, V., Wallace, D.C., and Sondheimer, N. (2017). Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. PLoS Comput Biol *13*.

Sorek, R., Ast, G., and Graur, D. (2002). Alu-Containing Exons are Alternatively Spliced. Genome Res *12*, 1060–1067.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology *98*, 503–517.

Speek, M. (2001). Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes. Mol Cell Biol *21*, 1973–1985.

Spielmann, M., Reichelt, G., Hertzberg, C., Trimborn, M., Mundlos, S., Horn, D., and Klopocki, E. (2011). Homozygous deletion of chromosome 15q13.3 including CHRNA7 causes severe mental retardation, seizures, muscular hypotonia, and the loss of KLF13 and TRPM1 potentially cause macrocytosis and congenital retinal dysfunction in siblings. European Journal of Medical Genetics *54*, e441–e445.

Srikanta, D., Sen, S.K., Huang, C.T., Conlin, E., Rhodes, R., and Batzer, M.A. (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. Genomics *93*, 205–212.

Stewart, J.B., and Chinnery, P.F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet *16*, 530–542.

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.-P., et al. (2011). A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. PLoS Genet *7*.

Suresh, B., Lee, J., Kim, H., and Ramakrishna, S. (2016). Regulation of pluripotency and differentiation by deubiquitinating enzymes. Cell Death Differ *23*, 1257–1264.

Szak, S.T., Pickeral, O.K., Landsman, D., and Boeke, J.D. (2003). Identifying related L1 retrotransposons by analyzing 3' transduced sequences. Genome Biol *4*, R30.

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. The American Journal of Human Genetics *101*, 700–715.

Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J., and Bahlo, M. (2018). Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. The American Journal of Human Genetics *103*, 858–873.

Taylor, R.W., and Turnbull, D.M. (2005). Mitochondrial DNA Mutations in Human Disease. Nat Rev Genet *6*, 389–402.

The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.-Y., Duan, J., Ophoff, R.A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. Nature Genetics *43*, 969–976.

Thevenon, J., Milh, M., Feillet, F., St-Onge, J., Duffourd, Y., Jugé, C., Roubertie, A., Héron, D., Mignot, C., Raffo, E., et al. (2014). Mutations in SLC13A5 Cause Autosomal-Recessive Epileptic Encephalopathy with Seizure Onset in the First Days of Life. Am J Hum Genet *95*, 113–120.

Thevenon, J., Duffourd, Y., Masurel-Paulet, A., Lefebvre, M., Feillet, F., Chehadeh-Djebbar, S.E., St-Onge, J., Steinmetz, A., Huet, F., Chouchane, M., et al. (2016). Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. Clinical Genetics *89*, 700–707.

Thung, D.T., de Ligt, J., Vissers, L.E., Steehouwer, M., Kroon, M., de Vries, P., Slagboom, E.P., Ye, K., Veltman, J.A., and Hehir-Kwa, J.Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol *15*.

Torene, R.I., Galens, K., Liu, S., Arvai, K., Borroto, C., Scuffins, J., Zhang, Z., Friedman, B., Sroka, H., Heeley, J., et al. (2020). Mobile element insertion detection in 89,874 clinical exomes. Genet Med 1–5.

Toydemir, R.M., Rutherford, A., Whitby, F.G., Jorde, L.B., Carey, J.C., and Bamshad, M.J. (2006). Mutations in embryonic myosin heavy chain (MYH3) cause Freeman-Sheldon syndrome and Sheldon-Hall syndrome. Nat Genet *38*, 561–565.

Tubio, J.M.C., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K., et al. (2014). Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M.I., Kidd, K.K., and Lenz, J. (2001). Insertional polymorphisms of full-length endogenous retroviruses in humans. Current Biology *11*, 1531–1535.

Uddin, M., Unda, B.K., Kwan, V., Holzapfel, N.T., White, S.H., Chalil, L., Woodbury-Smith, M., Ho, K.S., Harward, E., Murtaza, N., et al. (2018). OTUD7A Regulates Neurodevelopmental Phenotypes in the 15q13.3 Microdeletion Syndrome. The American Journal of Human Genetics *102*, 278–295.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3—new capabilities and interfaces. Nucleic Acids Res *40*, e115.

Vabres, P., Sorlin, A., Kholmanskikh, S.S., Demeer, B., St-Onge, J., Duffourd, Y., Kuentz, P., Courcet, J.-B., Carmignac, V., Garret, P., et al. (2019). Postzygotic inactivating mutations of RHOA cause a mosaic neuroectodermal syndrome. Nat Genet *51*, 1438–1441.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics *11*, 11.10.1-11.10.33.

Van Oven, M. (2015). PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Science International: Genetics Supplement Series *5*, e392–e394.

Van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. *30*, E386–E394.

Vasta, V., Ng, S.B., Turner, E.H., Shendure, J., and Hahn, S.H. (2009). Next generation sequence analysis for mitochondrial disorders. Genome Med *1*, 100.

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

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A *103*, 3220–3225.

Vissers, L.E.L.M. (2010). A de novo paradigm for mental retardation. Nature Genetics 6.

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.

Wagner, M., Berutti, R., Lorenz-Depiereux, B., Graf, E., Eckstein, G., Mayr, J.A., Meitinger, T., Ahting, U., Prokisch, H., Strom, T.M., et al. (2019). Mitochondrial DNA mutation analysis from exome sequencing—A more holistic approach in diagnostics of suspected mitochondrial disease. Journal of Inherited Metabolic Disease *42*, 909–917.

Wallace, D.C., Singh, G., Lott, M.T., Hodge, J.A., Schurr, T.G., Lezza, A.M., Elsas, L.J., and Nikoskelainen, E.K. (1988). Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. Science *242*, 1427–1430.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA Elements: A Hominid-specific Retroposon Family. Journal of Molecular Biology *354*, 994–1007.

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res *44*, W58–W63.

Wells, P.G., Bhatia, S., Drake, D.M., and Miller-Pinsler, L. (2016). Fetal oxidative stress mechanisms of neurodevelopmental deficits and exacerbation by ethanol and methamphetamine. Birth Defects Research Part C: Embryo Today: Reviews *108*, 108–130.

Wenger, A.M., Guturu, H., Bernstein, J.A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med *19*, 209–214.

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. Genome Res *15*, 1073–1078.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics *8*, 973–982.

Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet *385*, 1305–1314.

Wright, C.F., McRae, J.F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T.W., Jones, P., Prigmore, E., Rajan, D., Lord, J., et al. (2018). Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1133 families with developmental disorders. Genet Med *20*, 1216–1223.

Wu, J., Lee, W.-P., Ward, A., Walker, J.A., Konkel, M.K., Batzer, M.A., and Marth, G.T. (2014). Tangram: a comprehensive toolbox for mobile element insertion detection. BMC Genomics *15*.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., and Batzer, M.A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci U S A *103*, 17608–17613.

Xing, J., Kimura, H., Wang, C., Ishizuka, K., Kushima, I., Arioka, Y., Yoshimi, A., Nakamura, Y., Shiino, T., Oya-Ito, T., et al. (2016). Resequencing and Association Analysis of Six PSD-95-Related Genes as Possible Susceptibility Genes for Schizophrenia and Autism Spectrum Disorders. Sci Rep *6*.

Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N Engl J Med *369*, 1502–1511.

Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. JAMA *312*, 1870–1879.

Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., and Bruford, E.A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res *45*, D619–D625.

Ye, F., Samuels, D.C., Clark, T., and Guo, Y. (2014). High-Throughput Sequencing in Mitochondrial DNA Research. Mitochondrion *0*, 157–163.

Yin, J., Chen, W., Yang, H., Xue, M., and Schaaf, C.P. (2017). Chrna7 deficient mice manifest no consistent neuropsychiatric and behavioral phenotypes. Sci Rep *7*.

Yin, J., Chen, W., Chao, E.S., Soriano, S., Wang, L., Wang, W., Cummock, S.E., Tao, H., Pang, K., Liu, Z., et al. (2018). Otud7a Knockout Mice Recapitulate Many Neurological Features of 15q13.3 Microdeletion Syndrome. The American Journal of Human Genetics *102*, 296–308.

Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. Annu Rev Genomics Hum Genet *10*, 451–481.

Zhang, L., Bai, W., Yuan, N., and Du, Z. (2019). Comprehensively benchmarking applications for detecting copy number variation. PLOS Computational Biology *15*, e1007069.

Zhang, W., Cui, H., and Wong, L.-J.C. (2012). Comprehensive One-Step Molecular Analyses of Mitochondrial Genome by Massively Parallel Sequencing. Clinical Chemistry *58*, 1322–1331.

Zhao, F., Guan, M., Zhou, X., Yuan, M., Liang, M., Liu, Q., Liu, Y., Zhang, Y., Yang, L., Tong, Y., et al. (2009). Leber's hereditary optic neuropathy is associated with mitochondrial ND6 T14502C mutation. Biochemical and Biophysical Research Communications *389*, 466–472.

Zwieb, C., Van Nues, R.W., Rosenbald, M.A., Brown, J.D., and Samuelsson, T. (2005). A nomenclature for all signal recognition particle RNAs. RNA *11*, 7–13.

COMMUNICATIONS

Articles faisant l'objet de cette thèse

Philippine Garret, Celine Bris, Vincent Procaccio, Patrizia Amati Bonneau, Pierre Vabres, Nada Houcinat , Emilie Tisserant, François Feillet, Ange-Line Bruel, Virginie Quéré, Christophe Philippe, Arthur Sorlin, Frédéric Tran MauThem, Antonio Vitobello, JeanMarc Costa, Aïcha Boughalem, Detlef Trost, Laurence Faivre , Christel ThauvinRobinet, Yannis Duffourd. ***Deciphering exome sequencing data: bringing mitochondrial DNA variants to light.*** Hum Mutat, Dec 2019;40(12), 2430-2443

Philippine Garret, Frédéric Ebstein, Geoffroy Delplancq, Blandine Dozieres-Puyravel, Aïcha Boughalem, Stéphane Auvin, Yannis Duffourd, Sandro Klafack, Barbara A. Zieba, Sana Mahmoudi, Karun K. Singh, Laurence Duplomb, Christel Thauvin-Robinet, Jean-Marc Costa, Elke Krüger, Detlef Trost, Alain Verloes, Laurence Faivre, Antonio Vitobello. ***Report of the first patient with a homozygous OTUD7A variant responsible for epileptic encephalopathy and related proteasome dysfunction.*** Clin Genet, Apr 2020;07(4),567-575

Articles annexes

François Lecoquierre, Yannis Duffourd, Antonio Vitobello, Ange-Line Bruel, Benoit Urteaga, Christine Coubes, **Philippine Garret**, Sophie Nambot, Martin Chevarin, Thibaud Jouan, Sébastien Moutton, Orphanomix Physician's Group, Frédéric Tran-Mau-Them, Christophe Philippe, Arthur Sorlin, Laurence Faivre & Christel Thauvin-Robinet. ***Variant recurrence in neurodevelopmental disorders: the use of publicly available genomic data identifies clinically relevant pathogenic missense variants.*** GIM 2019 Apr

Ange-Line Bruel, Antonio Vitobello, Fred Tran Mau-Them, Sophie Nambot, Yannis Duffourd, Virginie Quéré, Paul Kuentz, **Philippine Garret**, Julien Thevenon, Sébastien Moutton, Daphné Lehalle, Nolwenn Jean-Marçais, Aurore Garde, Julian Delanne, Mathilde Lefebvre, François Lecoquierre, Detlef Trost, Megan Cho, Amber Begtrup, Aida Telegrafi, Orphanomix Physicians' Group, Pierre Vabres, Anne-Laure Mosca-Boidron, Patrick Callier, Christophe Philippe, Laurence Faivre, Christel Thauvin-Robinet. ***A 2.5 years' experience of GeneMatcher data-sharing: a powerful tool for identifying new genes responsible for rare diseases.*** Genet Med 2019 Jul

Pierre Vabres, Arthur Sorlin, Stanislav S Kholmanskikh, Bénédicte Demeer, Judith St-Onge, Yannis Duffourd, Paul Kuentz, Jean-Benoît Courcet, Virginie Carmignac, **Philippine Garret**, Didier Bessis, Odile Boute, Alain Bron, Guillaume Captier, Esther Carmi, Bernard Devauchelle, David Geneviève, Catherine Gondry-Jouet, Laurent Guibaud, Arnaud Lafon, Michèle Mathieu-Dramard, Julien Thevenon, William B Dobyns, Geneviève Bernard, Satyamaanasa Polubothu, Francesca Faravelli, Veronica A Kinsler, Christel Thauvin, Laurence Faivre, M Elizabeth Ross, Jean-Baptiste Rivière. ***Postzygotic inactivating mutations of RHOA cause a mosaic neuroectodermal syndrome.*** Nat Genet. 2019 Sep

Communications orales

Garret Philippine, Procaccio Vincent, Amati Bonneau Patrizia, Céline Bris, Vabres Pierre, Houcinat Nada, Tisserant Emilie, François Feillet, Bruel Ange-Line, Carmignac Virginie, Philippe Christophe, Sorlin Arthur, Tran Mau-Them Frédéric, Vitobello Antonio, Costa Jean-Marc, Boughalem Aïcha, Trost Detlef, Faivre Laurence, Thauvin-Robinet Christel, Duffourd Yannis. ***Décrypter les données de SHD d'exomes : les variations mitochondriales en lumière.*** 3ème jeudi de génétique – Paris – 18 octobre 2018

Garret Philippine, Procaccio Vincent, Amati Bonneau Patrizia, Céline Bris, Vabres Pierre, Houcinat Nada, Tisserant Emilie, François Feillet, Bruel Ange-Line, Carmignac Virginie, Philippe Christophe, Sorlin Arthur, Tran Mau-Them Frédéric, Vitobello Antonio, Costa Jean-Marc, Boughalem Aïcha, Trost Detlef, Faivre Laurence, Thauvin-Robinet Christel, Duffourd Yannis. ***Décrypter les données de SHD d'exomes : les variations mitochondriales en lumière.*** Forum des jeunes chercheurs – Dijon– 13 au 14 juin 2019

Garret Philippine, Procaccio Vincent, Amati Bonneau Patrizia, Céline Bris, Vabres Pierre, Houcinat Nada, Tisserant Emilie, François Feillet, Bruel Ange-Line, Carmignac Virginie, Philippe Christophe, Sorlin Arthur, Tran Mau-Them Frédéric, Vitobello Antonio, Costa Jean-Marc, Boughalem Aïcha, Trost Detlef, Faivre Laurence, Thauvin-Robinet Christel, Duffourd Yannis. ***Intérêt de la recherche de variations mitochondriales à partir de données de ES chez des patients atteints d'anomalies du développement et/ou de déficience intellectuelle.*** Club de Génétique de l'Est – Nancy– 28 novembre 2019

Garret Philippine, Procaccio Vincent, Amati Bonneau Patrizia, Céline Bris, Vabres Pierre, Houcinat Nada, Tisserant Emilie, François Feillet, Bruel Ange-Line, Carmignac Virginie, Philippe Christophe, Sorlin Arthur, Tran Mau-Them Frédéric, Vitobello Antonio, Costa Jean-Marc, Boughalem Aïcha, Trost Detlef, Faivre Laurence, Thauvin-Robinet Christel, Duffourd Yannis. ***Intérêt de la recherche de variations mitochondriales à partir de données de ES chez des patients atteints d'anomalies du développement et/ou de déficience intellectuelle.*** Assises de génétique humaine et médicale – Tours– 21 au 24 janvier 2020

Posters

P. Garret, V. Procaccio, P. Bonneau, P. Vabres, N. Houcinat, E. Tisserant, A.L. Bruel, V. Carmignac, C. Philippe, A. Sorlin, F. Tran Mau-Them, A. Vitobello, J.M. Costa, D. Trost, A. Boughalem, L. Faivre, C. Thauvin-Robinet, Y. Duffourd. ***Identification de variations pathogènes de l'ADN mitochondrial (ADNmt) à partir de données de séquençage à haut débit d'exome (SHD-E).*** Forum des jeunes chercheurs – Besançon – 15 au 16 juin 2018

P. Garret, V. Procaccio, P. Bonneau, P. Vabres, N. Houcinat, E. Tisserant, A.L. Bruel, V. Carmignac, P. Christophe, A. Sorlin, F. Tran Mau-Them, A. Vitobello, J.M. Costa, D. Trost, A. Boughalem, L. Faivre, C. Thauvin-Robinet, Y. Duffourd. ***Identification of mitochondrial disease variants on whole exome sequencing data.*** ASHG – San Diego, USA – 16 au 20 octobre 2018

P. Garret, C. Bris, V. Procaccio, P. Bonneau, P. Vabres, N. Houcinat, E. Tisserant, F. Feillet, A.L. Bruel, V. Quéré, C. Philippe, A. Sorlin, F. Tran Mau-Them, A. Vitobello, J.M. Costa, A. Boughalem, D. Trost, L. Faivre, C. Thauvin-Robinet, Y. Duffourd. ***Deciphering exome sequencing data: bringing mitochondrial DNA variants to light.*** ESHG – Gothenburg, Sweden – 15 au 18 juin 2019

Garret Philippine, Delplancq Geoffroy, Ebstein Frédéric, Dozieres-Puyravel Blandine, Boughalem Aïcha, Auvin Stéphane, Duffourd Yannis, Zieba Barbara, Mahmoudi Sana, Singh Karun, Thauvin-Robinet Christel, Costa XJean-Marc, Krüger Elke,Trost Detlef, Verloes Alain, Faivre Laurence, Vitobello Antonio. ***Report of the first patient with a homozygous OTUD7A variant responsible for epileptic encephalopathy.*** ASHG – Houston, USA – 15 au 19 octobre 2019

Garret Philippine, Delplancq Geoffroy, Ebstein Frédéric, Dozieres-Puyravel Blandine, Boughalem Aïcha, Auvin Stéphane, Duffourd Yannis, Zieba Barbara, Mahmoudi Sana, Singh Karun, Thauvin-Robinet Christel, Costa Jean-Marc, Krüger Elke, Trost Detlef, Verloes Alain, Faivre Laurence, Vitobello Antonio. ***Description du premier patient atteint d'encéphalopathie épileptique porteur d'une variation homozygote dans le gène OTUD7A.*** Assises de génétique humaine et médicale – Tours– 21 au 24 janvier 2020

Garret Philippine, Tisserant Emilie, Boughalem Aïcha, Costa Jean-Marc, Trost Detlef, Vitobello Antonio, Faivre Laurence, Thauvin-Robinet Christel, Duffourd Yannis. ***Identification de l'insertion d'éléments mobiles dans des régions géniques à partir de données de séquençage à haut débit d'exome et de génome chez des patients atteints d'anomalies du développement et/ou de déficience intellectuelle.*** Assises de génétique humaine et médicale – Tours– 21 au 24 janvier 2020

ÉCOLE DOCTORALE Environnements - Santé Bourgogne | Franche-Comté

Titre : Approches bioinformatiques innovantes pour l'analyse de données de séquençage à haut-débit appliquées à l'étude de pathologies génétiques rares avec anomalies du développement

Mots clés : bioinformatique - exome - maladies génétiques rares

Résumé : L'avènement du séquençage haut débit d'exome (ES) en diagnostic et en recherche ces dernières années a conduit à l'identification des bases génétiques de nombreuses pathologies mendéliennes, permettant de résoudre de nombreuses situations d'errance diagnostique. Néanmoins, l'analyse des données de ES permet uniquement d'identifier des variations pathogènes ou probablement pathogènes dans 30 à 45 % des situations sans diagnostic. En effet, certaines limites existent, tant au niveau clinique, moléculaire et bioinformatique. L'évolution constante des connaissances cliniques, du nombre de nouveaux gènes impliqués en pathologie humaine, et des corrélations clinico-biologique a un impact important sur l'analyse des données, entraînant une amélioration progressive de la recherche diagnostique. Des limites techniques inhérentes à la technologie, avec en particulier des régions non couvertes, existent, mais se sont également significativement réduites ces dernières années. Enfin, au-delà de l'analyse de SNV et de CNV, d'autres anomalies génétiques peuvent être responsables de maladies rares, nécessitant un développement bioinformatique pour optimiser les résultats.

Bien que le séquençage à haut débit du génome permette de résoudre des observations, en particulier en cas de variations dans les régions non codantes ou les variants de structure, il existe encore de nombreuses informations à extraire et à exploiter à partir des données de ES.

L'objectif de cette thèse a donc été de participer à l'amélioration des approches bioinformatiques d'analyse de données de ES pour l'identification de nouveaux gènes ou mécanismes moléculaires impliqués dans des maladies génétiques rares afin de réduire l'errance diagnostique des patients.

Plusieurs stratégies ont ainsi été mises en place. La première stratégie a consisté en une réanalyse recherche de données de 80 patients ayant bénéficié d'un ES au laboratoire CERBA (thèse CIFRE) dont la lecture diagnostique était négative. Elle a conduit à la mise en évidence deux nouveaux gènes candidats dans la déficience intellectuelle syndromique, dont le gène *OTUD7A* (article 1). La deuxième stratégie a consisté en la mise au point d'un pipeline bioinformatique pour extraire les données du génome mitochondrial à partir des données de ES. L'ADN mitochondrial n'est pas ciblé par les kits de capture d'exome mais peut être extrait des données capturées indirectement, rendant son analyse possible à partir de données de ES préexistantes. A partir de la collection GAD d'exomes de patients sans diagnostic, deux variations causales ont été identifiées chez deux individus atteints de troubles neuro-développementaux sur 928 personnes étudiées, et ainsi résoudre une errance diagnostique dans 0,2 % des patients sans diagnostic (article 2). La troisième stratégie a consisté en la mise en place d'un pipeline bioinformatique d'identification des éléments mobiles au sein des données d'exome, étant attendu qu'environ 0,3 % des variations pathogènes du génome humain ont pour origine l'insertion *de novo* d'un élément mobile. A partir de la collection GAD d'exomes de 3322 individus (2500 cas index) sans diagnostic, cette étape a permis d'identifier deux cas solides en lien avec l'insertion d'un élément Alu au sein d'un exon du gène *FERMT1* et du gène *GRIN2B* (article 3 en cours d'écriture).

Cette thèse a permis de repousser certaines limites de la technologie d'exome. D'autres perspectives existent, et sont explorées par l'équipe, en lien avec le projet Européen Solve-RD.

Title : Innovative bioinformatics approaches for the analysis of high-throughput sequencing data applied to the study of rare genetic pathologies with developmental abnormalities

Keywords : bioinformatics - exome - rare genetic diseases

Abstract : In recent years, the advent of exome sequencing (ES) in diagnosis and research has led to the identification of the genetic bases of many Mendelian disorders. In turn, this has allowed researchers to resolve many undiagnosed cases. Nevertheless, ES data analysis only leads to the identification of pathogenic or likely pathogenic variants in 30 to 45 % of these unsolved cases, which are known as diagnostic odysseys. Indeed, There are some limitations at the clinical, molecular and bioinformatics levels. The constant progression of clinical knowledge, of the number of genes involved in human diseases, and of clinical-biological correlations have had a significant impact on data analysis, leading to a progressive improvement in diagnostic research. However there are limits to the current technologies, particularly for regions that are not covered, though these limits have been significantly reduced over the last few years. Although genome sequencing will certainly resolve a number of undiagnosed cases, especially in case of non-coding or structural variants, there is still a lot information to be extracted and analyzed from ES data. Finally, beyond SNV and CNV analyses, other genetic events can be involved in rare disorders, requiring the development of bioinformatics to optimize results.

The aim of this project was therefore to improve bioinformatics approaches to ES data analysis in order to identify new molecular mechanisms involved in rare genetic disorders and thus reduce the number of diagnostic odysseys.

Several strategies were established. The first consisted in reanalyzing ES data from 80 undiagnosed patients who were sequenced by the Laboratoire CERBA (CIFRE thesis). This led to the identification of 2 new candidate genes involved in ID, particularly the *OTUD7A* gene (Article 1). The second strategy was to develop a bioinformatics pipeline in order to extract mitochondrial DNA data from ES data. The mitochondrial genome is not targeted by exome capture kits but can be extracted from off-target data, which provides an opportunity to analyze it from preexisting ES data. From the GAD exomes cohort of undiagnosed patients, 2 causal variations were identified in 2 out of 928 individuals with a neuro-developmental disorder. This approach therefore resolved the diagnostic odyssey of 0.2 % of undiagnosed patients (Article 2). The third strategy was to develop a bioinformatics pipeline to identify the insertion of mobile elements within ES data, with the expectation that about 0.03 % of pathogenic variants originate from *de novo* insertion of mobile elements. This step lead to the identification of 2 cases of Alu element insertion in *FERMT1* and *GRINB2* gene exons from the GAD exomes cohort of 3322 undiagnosed individuals (2500 probands) (Article 3, in progress).

The research undertaken for this PhD has expanded some of the limits of ES. Still other perspectives exist and are currently being explored by the GAD team in collaboration with the European Solve-RD project.

UBFC UNIVERSITE BOURGOGNE FRANCHE-COMTE

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon