Université de Montréal

Identification de gènes impliqués dans les ataxies épisodiques par combinaison de séquençages génomique et transcriptomique

Par

Sébastien Audet

Département de Neurosciences, Université de Montréal, Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade Maîtrise ès Sciences (M.Sc.) en Neurosciences

Décembre 2021

Université de Montréal

Unité académique : Département Neurosciences, Université de Montréal, Faculté de Médecine

---

Ce mémoire intitulé

Identification de gènes impliqués dans les ataxies épisodiques par combinaison de séquençages génomique et transcriptomique

Présenté par

Sébastien Audet

A été évalué(e) par un jury composé des personnes suivantes

Elsa Rossignol
Président-rapporteur

Martine Tétreault
Directeur de recherche

Mark E. Samuels
Membre du jury

Résumé

Cette étude pilote vise à développer une méthode d'analyse intégrative qui permet d'augmenter le taux de réussite du diagnostic clinique des mutations génétiques rares. De plus, l'identification de nouveaux gènes associés à l'ataxie épisodique (EA) et l'évaluation de nouveaux algorithmes de prédiction, pour un examen de variants plus robuste, découleront de l'enquête.

Caractérisé par une perte sporadique de la coordination des mouvements volontaires, l'EA se manifeste généralement tardivement, avec une hétérogénéité clinique et génétique élevée, compliquant largement l'obtention d'un diagnostic précis. Alors que quatre gènes ont été liés aux huit sous-types d'EA, de nombreux patients demeurent sans diagnostic moléculaire dû aux limites des méthodes de séquençage d'ADN. Ces lacunes accentuent l'intérêt d'implanter le séquençage de l'ARN en milieu clinique, afin d'obtenir l'information fonctionnelle offerte par l'approche.

Des patients atteints d'EA, sans diagnostic moléculaire malgré un examen approfondi, ont été recrutés à Montréal. Le séquençage du génome entier (WGS) et de l'ARN a été effectué sur des échantillons de sang pour identifier les variants nucléotidiques, l'expression différentielle, les événements d'épissage ainsi que les expansions de microsatellites. Plusieurs algorithmes de prédiction de la pathogénicité récents ont été choisis pour être testés parallèlement aux algorithmes standard. Des données WGS provenant d'un trio familial atteint de pathologies neurologiques ont également été soumises au pipeline génomique développé pour la cohorte EA.

Des variants candidats ont été identifiés pour chaque patient en fonction des scores de pathogénicité, de la rareté des événements génétiques et des informations fonctionnelles et cliniques connues pour un gène altéré donné. Parmi les découvertes figurent des mutations non-sens, des faux-sens, de l'épissage alternatif ainsi que des expansions nucléotidiques dans des gènes associés aux ataxies spinocérébelleuses ou aux paraplégies spastiques. En plus d'être présents dans les ensembles de données de séquençage disponibles pour chaque patient, les événements génomiques ont été vérifiés par séquençage Sanger de l'ADN et de l'ARN lorsque possible. Les effets fonctionnels potentiels, prédits principalement à partir du RNA-seq et suggérant une expression anormale de l'ARNm, ont également été évalués par amplification PCR

et qPCR traditionnelle. À ce jour, quatre des dix patients ont reçu ou sont en voie de recevoir un diagnostic clinique, et quatre autres présentent d'excellents candidats moléculaires pour expliquer une pathologie ataxique.

Ce projet devrait permettre un diagnostic mieux défini, conduisant à une meilleure qualité de vie, une meilleure évaluation du pronostic et une meilleure prise en charge des patients. L'identification de modulateurs génétiques chez certains d'entre eux devrait également permettre une meilleure caractérisation clinique des conditions rapportées, bénéficiant les évaluations symptomatiques futures. De plus, la méta-analyse des données RNA-seq offre le potentiel de découvrir des régulateurs de pathogenèse communs à l'EA. Il favorisera également l'approche intégrative pour un plus large éventail de troubles et pourrait éventuellement conduire à de nouvelles stratégies thérapeutiques.

Mots-clés : Génomique, transcriptomique, WGS, RNA-seq, génétique clinique, ataxies, bio-informatique, SpliceAI, ExpansionHunter, analyse intégrative.

Abstract

This pilot study aims to develop an integrative analysis method that allows for an increased diagnosis success rate of rare genetic mutations. Moreover, identification of novel genes associated with Episodic Ataxia (EA) and evaluation of new AI-generated prediction algorithms, for a more robust variant examination, will ensue from the investigation.

Characterized by sporadic loss of voluntary movement coordination, EA typically manifest with a late onset as well as high-clinical and genetic heterogeneity, setting additional hurdles to diagnosis. While four genes have been linked to the eight subtypes of EA, many patients are left without molecular diagnosis due to the limitations of individual DNA-sequencing methods, which can be mitigated by the functional overview that RNA sequencing (RNA-seq) offers.

EA patients, lacking molecular diagnosis despite in-depth examination, were recruited in Montreal. Whole-Genome sequencing (WGS) and RNA-seq were performed on blood samples to identify single nucleotide variants, differential expression, splicing events, structural variants and repeat expansions. Multiple recent pathogenicity prediction algorithms were chosen for testing concurrently to standard ones, in order to evaluate their performance and potential for clinical pipelines integration. WGS data of a family trio from France, in which the father and the daughter present neurologic pathologies, were also processed through the genomic pipeline that was developed for the EA cohort in order to identify the cause of their disorder.

Candidate variants were identified for each patient according to pathogenicity scores, rarity of genetic events, and known functional as well as clinical information for a given altered gene. Among the findings are truncations, missenses, alternative splicing, and repeat expansions in genes already associated to either spinocerebellar ataxia or spastic paraplegia. In addition to being present in both datasets when available, validation of these interesting genomic events has been performed through Sanger Sequencing of both DNA and RNA when feasible. For strong candidates where the available functional information from RNA-seq suggests abnormal mRNA expression, validation includes PCR amplification as well as a traditional qPCR to support effects on transcripts. To this day, four out of ten patients have received or are on the verge of receiving a diagnosis, and four others are carrying excellent molecular candidates requiring further validation to explain their ataxic pathologies.

This project should provide more defined diagnosis, leading to better quality of life, better evaluation of prognosis and better management of care for patients. Identification of genetic modifier in some of them should also allow for a better clinical characterization of the reported conditions, benefiting future patient examinations. A meta-analysis of our patients' transcriptomic profiles could also uncover commonly affected pathways in EA development. It will also promote the integrative approach for a larger spectrum of disorders and might eventually lead to new therapeutic strategies.

Keywords: Genomic, transcriptomic, WGS, RNA-seq, clinical genetics, ataxia, bioinformatics, SpliceAI, ExpansionHunter, integrative analysis.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ADN : Acides désoxyribonucléiques

ADNc : ADN complémentaire

ACMG : Collège américain de génétique médicale

AD : Autosomique dominant

AG : Gain de site accepteur

AL : Perte de site accepteur

AF : Fréquence allélique

AR : Autosomique récessif

ARCA : Ataxie cérébelleuse autosomique récessive

ARN : Acides ribonucléiques

AS : Épissage alternatif

AVEC : Ataxie avec déficience en vitamine E

BWA : Aligneur de Burrow-Wheeler

CANVAS : Syndrome d'ataxie cérébelleuse, neuropathie et aréflexie vestibulaire

CAPOS : Syndrome d'ataxie cérébelleuse, aréflexie, pieds creux, atrophie optique et perte auditive neurosensorielle

CESGQ : Centre d'expertise et de services Génome Québec

CHUM : Centre hospitalier de l'Université de Montréal

DG : Gain de site donneur

DL : Perte de site donneur

DMSO : Diméthyle sulfoxyde

EA : Ataxie épisodique

EDTA : Acide éthylènediaminetétraacétique

EH : ExpansionHunter

EHdn : ExpansionHunter Denovo

FBS : Sérum fœtal bovin

FRDA : Ataxie de Friedreich

GATK : Boîte à outils d'analyse du génome

indels : Insertions et délétions de petite taille

NCBI : Centre national des biotechnologies de l'information

NMD : Dégradation médiée par ubiquitination

pb : Paires de bases

PCR : Réaction en chaîne par polymérase

pLI : Score de résistance à la perte de fonction

qPCR : PCR quantitative

RNA-seq : Séquençage de l'ARN

SCA : Ataxie spinocérébelleuse

SCAR : Ataxie spinocérébelleuse autosomique récessive

SNV : Variant nucléotidique

STR : Répétitions en tandem court

SV : Variant structurel

UTR : Régions non traduites de l'ARNm

VUS : Variant de signification inconnue

WGS : Séquençage du génome complet

À toi qui veilles maintenant sur moi de là-haut, merci de m'avoir poussé à constamment chercher le meilleur de moi-même, et d'avoir cru en moi.

Remerciements

Je tiens premièrement à remercier ma directrice de recherche, Martine Tétreault, pour son support continu au courant des deux dernières années. Pour la confiance qu'elle a placée en moi malgré mes doutes personnels, je n'aurais pas pu demander une meilleure superviseure. Merci d'avoir guidé mon parcours avec toutes tes connaissances et ta sagesse, en plus de mettre en place un laboratoire où l'ambiance est toujours positive et prône à une bonne recherche. Grâce à toi, j'ai réellement l'impression d'être devenu un meilleur chercheur, et avec un peu de chance, une meilleure personne.

Un gros merci à tous les autres membres du laboratoire, qui ont su rendre ma maîtrise mémorable. À Jennifer pour les conversations épanchées lors des longues soirées de travail, à Camille pour ses connaissances lors de discussions rétrospectives sur des expériences ratées, à Valérie pour sa capacité à m'écouter me plaindre fréquemment sur la bio-informatique, et à Jean pour sa simple camaraderie. Je n'oublie pas tous mes autres collègues de laboratoire, qui je suis sûr auront beaucoup de succès dans leurs futures carrières : Camberly, Lovatiana, Gaël, Nab, Marjorie, Jade, Adrien, Annie, Éric, Renaud, Moustafa.

Je remercie également les autres membres du CHUM qui ont appuyé d'une façon ou d'une autre mon parcours. Quelques pensées particulières pour Julie, dont le savoir et la bonne humeur ont fréquemment sauvé la situation, Romane pour son mentorat et la confiance qu'elle a eue en moi, et Floriane pour sa présence d'esprit ainsi que son calme réconfortant lors de l'optimisation de la mise en place du séquençage Nanopore d'un projet secondaire.

Finalement, un gros merci à ma famille et mes amis pour leur support inconditionnel malgré la visible confusion lors des explications de la nature de mon travail. Particulièrement toi Kelly, pour ta patience, ta compassion, ainsi que pour ton amour.

1 - Introduction

1.1 Maladies génétiques rares

Malgré le fait qu'ils soient caractérisés par une faible fréquence individuelle dans la population, l'ensemble combiné des maladies rares représente en fait un énorme fardeau médical (1). En effet, selon la base de données Orphanet, c'est 6172 maladies rares différentes qui affectent cumulativement 3.5-5.9% de la population mondiale, soit approximativement 263 à 446 millions de personnes (2). La majorité de ces pathologies sont d'origine génétique, et affecte fréquemment un seul locus chromosomique, les caractérisant plus précisément de désordres mendéliens. Plusieurs informations sont pertinentes à la définition d'un pedigree génétique : le mode de transmissions indique si le gène mutant est présent sur un gonosome, autrement connu comme chromosome sexuel, ou bien sur un autosome. Il est question de transmission dominante (AD) lorsqu'un seul allèle affecté cause l'apparition de phénotypes, alors que la nécessité de porter deux copies du gène mutant caractérise un désordre autosomique récessif (AR). Pour ce dernier cas, cela signifie qu'une partie de la population est porteur sain du variant en question. En effet, des milliers de mutations germinales s'accumulent dans le génome humain lors du développement zygotique (3), la grande majorité étant cependant non-codante, puisque la proportion d'ADN codant représente moins de 2% du génome total (4). Une mutation est dite *de novo* lorsqu'elle résulte de la recombinaison germinale. Tout cela complique l'identification et l'interprétation des mutations pathogéniques, puisqu'il est impossible d'investiguer l'impact de la variabilité génétique de chaque nucléotide du génome. À cet effet, la recherche de motifs fonctionnels récurrents et conservés aide à la prédiction d'impact pathogénique d'un variant, et par le fait même à la classification de l'intérêt des mutations non documentées (5). Un autre aspect qui complique la définition d'une cause pathogénique est la possibilité d'un hétérozygote composé, où un locus est dysfonctionnel dû à deux variants d'origine distincte (6).

Il existe plusieurs types de mutations pouvant affecter la structure ou la fonction d'un gène : alors qu'un variant nucléotidique (SNV) synonyme indique que l'altération de base azotée ne modifie pas un codon protéique correspondant, les SNV non synonymes induisent un changement d'acide

aminé (faux-sens) ou l'introduction d'un nouveau codon stop (non-sens). Dans un cas où un SNV, codant ou intronique, modifie le patron d'épissage canonique d'un gène, il est question d'épissage alternatif (AS). Les variants structurels (SV) regroupent arbitrairement les modifications supérieures à 1Kb, soient les déséquilibres génomiques (insertions/délétions), translocations, inversions, et variants de nombre de copies (CNV). Les insertions et délétions de petite taille se nomment quant à eux indels (7). Finalement, les expansions nucléotidiques réfèrent à la réplication d'un microsatellite de quelques nucléotides. La localisation des mutations joue également un rôle important sur son potentiel pathogénique. Celles-ci peuvent être dans une région transcrite exonique ou intronique, dans une région non codante, ou encore dans une région régulatrice telle qu'une séquence promotrice, modulatrices, ou bien correspondant à un ARN non codant. Ajoutant à la complexité du génome humain, l'empreinte épigénétique contribue fortement à la régulation de l'expression de nombreux gènes. La modification de patrons précis de méthylation ou encore d'acétylation peut donc avoir un impact majeur sur des voies moléculaires essentielles aux fonctions cellulaires de l'humain (8).

1.2 Approche clinique diagnostique

Suite à une évaluation clinique généralement exhaustive afin de définir la présentation phénotypique d'un désordre, il est généralement possible d'attribuer relativement précisément la bonne classe de pathologie au patient, mais pas nécessairement de définir le type exact. À cet effet, seule une investigation génétique permet l'obtention d'un diagnostic moléculaire officiel. Généralement, les cliniciens ont recours à des tests de séquençage nouvelle génération (NGS) par panels, puisque ceux-ci regroupent tous les gènes causatifs connus pour une pathologie donnée (9). Les tests ciblant un ou plusieurs gènes précis sont parfois considérés (10). Plusieurs autres informations sont utilisées pour définir l'approche moléculaire pertinente, telles que l'observation de phénotypes caractéristiques à une maladie spécifique, le mode de transmission suspecté, l'ethnicité du patient pouvant affecter la fréquence pathogénique de certains gènes, ou encore la possibilité de traitement pour certaines causes hypothétiques. Par exemple, pour un patient qui présenterait des symptômes précoces de dégénération neuromusculaire progressive, plusieurs compagnies commerciales offrent des panels couvrant les 35 gènes possiblement impliqués dans le développement de la sclérose latérale amyotrophique (11). Cette méthode est

relativement efficace et s'améliore constamment avec la nouvelle littérature, en plus de permettre de réduire le coût ainsi que la charge de travail à un minimum pour l'obtention d'un diagnostic génétique. Cependant, le test est généralement restreint à l'information obtenue aux régions exoniques de seulement quelques gènes, en plus de n'offrir un diagnostic moléculaire que lorsque le variant identifié est préalablement classifié comme pathogénique ou probablement pathogénique selon les critères officiels du collège américain de génétique médicale (ACMG) (12). En fonction de ceux-ci, la majorité des mutations génétiques rares sont classifiées comme variants de signification inconnue (VUS) jusqu'à ce qu'il y ait une validation expérimentale de l'effet de la mutation sur la fonction du gène. Dans certains cas, le séquençage d'exome, également restreint aux régions codantes, ou encore le séquençage « génome rapide » sont offerts par certains hôpitaux tels que Sainte-Justine (13, 14), mais il demeure évident que la recherche fondamentale est une composante essentielle à l'interprétation clinique (15). Habituellement, pour ces cas plus complexes, davantage de ressources peuvent être déployées afin d'identifier les meilleurs candidats via une analyse poussée de la littérature disponible sur les fonctions, l'expression, et les interactions d'un gène donné. Plusieurs outils de prédiction *in silico* sont également utiles à cette évaluation.

1.3 Ataxies

Dans le cadre de ce mémoire, il sera principalement question de maladies neurodégénératives où l'examen clinique suggère fortement des ataxies. Cette classe de maladies est caractérisée par des atteintes motrices et proprioceptives, principalement la coordination de mouvements volontaires, habituellement causées par une dégénérescence progressive des neurones du cervelet ou par l'atteinte de structures adjacentes (16).

1.3.1 Cervelet

1.3.1.1 Structure et fonctions

Les conséquences cliniques de la dégénération neuronale varient énormément en fonction de la structure affectée. Le cervelet, étant à la base du cerveau, est connu comme ayant un rôle majeur dans le contrôle moteur en plus de participer à plusieurs fonctions cognitives telles que le langage

et la mémoire de travail (17). Structurellement, le cervelet s'apparente au cortex cérébral, étant séparé en deux hémisphères ainsi que trois lobes principaux, soient flocculo-nodulaire, antérieur et postérieur. Au niveau cellulaire, les neurones de Purkinje et les cellules granulaires sont les constituants centraux du circuit neuronal fonctionnel complexe (Figure 1), qui implique aussi des signaux axonaux reliant les noyaux cérébelleux profonds et des entrées de l'extérieur du cervelet. Les transmissions neuronales permettant les différentes fonctions cérébelleuses peuvent également être modulées par diverses entrées dopaminergiques, noradrénalinergiques, cholinergiques et sérotoninergiques affectant principalement l'efficacité de l'apprentissage neuronal (18). À cet effet, il est intéressant de noter que la plasticité des neurones de Purkinje a été démontrée expérimentalement, et joue un rôle important dans les capacités motrices de l'être humain (19).

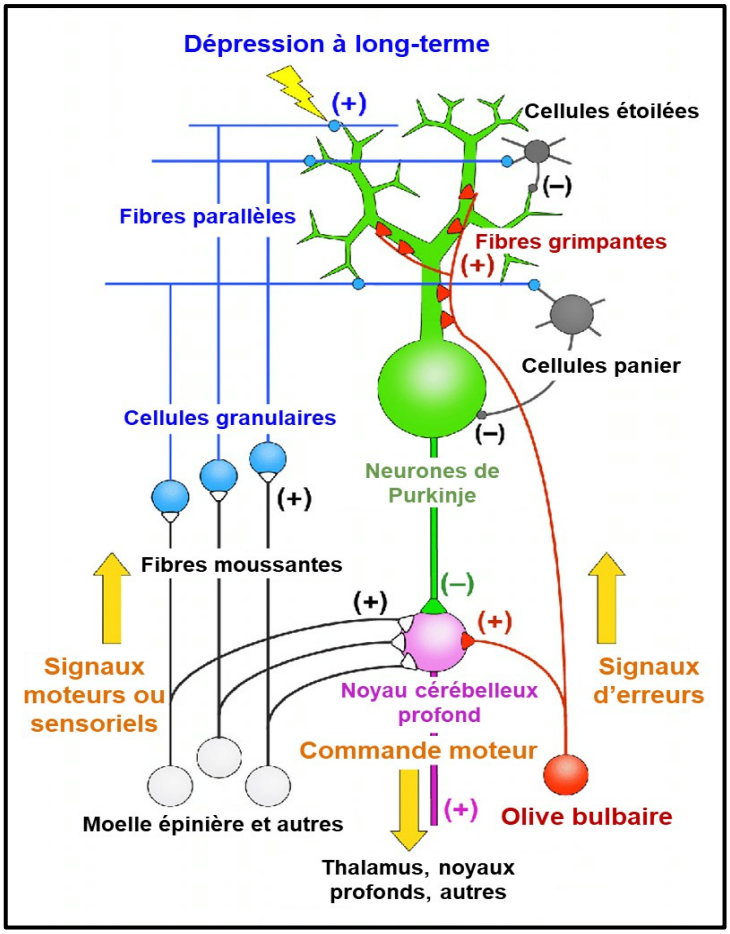

Figure 1. Représentation simplifiée du circuit synaptique du cervelet. Schéma détaillé des signaux du circuit neuronal (20). Traduction française tirée de l'article de Kano M., *et al*. (2017).

1.3.1.2 Pathologies

Il existe une multitude de maladies impliquant le dysfonctionnement des circuits cérébelleux, la majorité étant bien sûr associée à des symptômes moteurs tels que le manque de coordination au niveau de la balance, de la démarche, des mouvements fins (dysmétrie), de la parole (dysarthrie) ainsi que des mouvements de l'œil. Cependant, certaines études suggèrent que des désordres purement psychiatriques, tels que le syndrome cognitivo-affectif du cervelet, peuvent découler d'une atteinte des fonctions cognitives cérébelleuses (21). La classification des maladies du cervelet se rattache principalement à la source de l'atteinte neuronale, soit primaire suite à une mutation génétique, ou secondaire suite à diverses pathologies. Il est question de désordre d'origine toxique lorsqu'il y a ingestion involontaire ou excessive de molécules neurotoxiques telles que du mercure, de l'alcool, ou encore du monoxyde de carbone. La mort des neurones cérébelleux peut également survenir suite à une réaction auto-immune inflammatoire comme dans la sclérose en plaques (11), à un accident vasculaire tel qu'une ischémie ou hémorragie (22), à une infection virale ou bactérienne comme dans la maladie de Lyme (23), ou encore suite à un traumatisme crânien important (24). Quoique les cancers ont un aspect génétique fréquemment important, l'impact sur les fonctions cérébelleuses est majoritairement relié à la formation de tumeur, et donc les mutations pathogéniques ont un effet indirect sur les phénotypes ataxiques. De façon similaire, les altérations du système nerveux ne sont que l'une des conséquences des désordres métaboliques comme les déficiences en vitamine E (AVED) (25). Les troubles neurodéveloppementaux ont un impact qui n'est pas restreint au cervelet, mais l'association causative de mutations a été réalisée avec de nombreux gènes neuronaux. Néanmoins, les symptômes d'ataxies sont dits secondaires aux anormalités structurelles de la formation du cervelet. Finalement, de nombreuses maladies génétiques rares affectent particulièrement la structure et les fonctions du cervelet, et cela de façon progressive. Il est alors question d'ataxies primaires, où la neurodégénération peut être d'origine héréditaire ou sporadique (16).

1.3.2 Ataxies autosomiques récessives

Une ataxie cérébelleuse est dite autosomique récessive (ARCA) lorsque des mutations sont nécessaires sur les deux allèles pour provoquer des phénotypes. Comme la perte des deux copies d'un gène a tendance à avoir un impact moléculaire plus important sur les voies fonctionnelles

associées, cela conduit généralement à une présentation clinique plus précoce et sévère. La forme la plus connue de ce type de pathologie est l'ataxie de Friedreich (FRDA), où la cause classique est une expansion bi-allélique du microsatellite CAG dans l'intron 1 de la frataxine (*FXN*) au-dessus du seuil pathogénique de 34 à 65 répétitions (26). Le diagnostic d'ataxie progressive survient généralement entre 10 à 15 ans, et se caractérise typiquement par une dysarthrie importante, une faiblesse et spasticité musculaire, en plus de certains symptômes extracérébelleux tels qu'une scoliose, une dysfonction rénale, une cardiomyopathie ou encore un diabète de type 1. Parmi les autres formes récessives, les ataxies spinocérébelleuses autosomiques récessives (SCAR) sont proéminentes puisqu'il s'agit de la nomenclature à utiliser pour une forme typique : il existe au moins 28 types de SCAR où le gène en cause diffère, mais la présentation clinique a toujours une composante ataxique primaire et une apparition avant l'âge adulte, tel qu'observé avec FRDA (27). L'atrophie du cervelet est omniprésente, s'étendant parfois à d'autres structures du cerveau, et le pronostic le plus fréquent est une perte progressive des capacités motrices atteignant éventuellement un stade débilitant. Au total, plus de 60 formes d'ARCA ont été décrites avec ces caractéristiques phénotypiques en premier plan (28). On retrouve notamment l'ataxie récessive spastique Charlevoix-Saguenay (ARSACS), largement étudié pour sa prévalence amplifiée par un effet fondateur, et les ataxies spastiques (SPAX), tous deux caractérisées par une forte composante spastique conjointement à l'ataxie (29). Il est important de noter qu'en plus d'impliquer un très grand nombre de gènes, ceux-ci peuvent causer des désordres de forme dominante parallèlement à leur transmission récessive, dépendamment du type de variant en cause. Un exemple de ce phénomène est visible chez le gène du récepteur inositol 1,4,5-trisphosphate 1 (*ITPR1*) qui, en fonction de l'emplacement et du type de variant, peut causer une ataxie spinocérébelleuse (SCA) de type 15 (AD) ou 29 (AD/AR) (30).

1.3.3 Ataxies autosomiques dominantes

La majorité des formes AD sont classifiées comme SCA, fréquemment caractérisées par une apparition tardive et progressive, une démarche ataxique, une dysmétrie, une dysarthrie, ainsi qu'une atrophie du cervelet (16). Il existe à ce jour près de 50 sous-types de SCA ayant des gènes causatifs associés, il est donc quasiment impossible de différencier précisément les formes d'un point de vue clinique, dû au grand chevauchement phénotypique. De plus, une grande variabilité

de pénétrance peut être observée d'un patient à l'autre, parfois même intrafamiliale. Dans certains cas, le phénomène est explicable via l'effet d'anticipation, où un plus grand nombre de répétitions du microsatellite toxique augmente la gravité de l'impact moléculaire (31). À cet effet, plusieurs gènes causant des SCA portent fréquemment des expansions nucléotidiques, soient les ataxines (ATXN) *ATXN1*, *ATXN2*, *ATXN3*, *ATXN7*, *ATXN8OS*, *ATXN10*, ainsi que *TBP*, *PPP2R2B*, *NOP56* et *CACNA1A* (16, 32). Le gène de l'atrophine (*ATN1*) est lui aussi caractérisé par une répétition causant une forme dominante d'ataxie cérébelleuse où les phénotypes plus complexes s'apparentent à la chorée de Huntington. Un fait remarquable est l'importance indéniable des protéines ataxines aux diverses fonctions du cervelet, elles qui jouent fréquemment différents rôles de régulation par liaison protéique (33, 34). Un autre aspect intéressant est la capacité de la sous-unité alpha 1A du canal calcique voltage-dépendant (*CACNA1A*) à causer parallèlement ou simultanément une SCA de type 6, une migraine hémiplégique familiale de type 1, une épilepsie encéphalopathique avec troubles développementaux, ou une ataxie épisodique (EA) de type 2 en fonction de la localisation et de la nature de la mutation *CACNA1A* d'un patient (35, 36). Cela souligne une fois de plus la très grande variabilité phénotypique reliée aux ataxies.

1.3.4 Ataxies épisodiques

Les EA représentent une forme relativement unique d'ataxie, où la présentation clinique est très similaire à celle des SCA, à l'exception d'une manifestation seulement sporadique des symptômes ataxiques (37). En effet, ceux-ci ne sont généralement visibles que suite à la présence d'un déclencheur symptomatique, et la fréquence des épisodes peut varier de quelques attaques par années à plusieurs crises par jour. Divers facteurs environnementaux sont en mesure d'induire ces épisodes, tels que l'exercice physique, un stress émotionnel, ou encore la simple consommation de caféine ou de médicaments (38). La fréquence des attaques symptomatiques augmente avec la neurodégénérescence, et il arrive que l'ataxie persiste entre les crises pour certains patients, allant parfois jusqu'à évoluer en une forme s'appariant davantage à une SCA. C'est un fait particulier qui est important à considérer : un diagnostic d'EA est difficile à émettre sans identification de cause moléculaire, puisque la caractéristique épisodique peut disparaître avec la progression pathogénique. Il est donc important de ne pas mettre de côté la possibilité d'un gène typique d'une autre forme d'ataxie chez un patient cliniquement atypique.

1.3.4.1 Présentation clinique

Outre l'irrégularité de la présentation phénotypique, plusieurs symptômes caractérisent habituellement chacun des huit sous-types connus d'ataxie épisodique. La présentation typique est fréquemment tardive avec une progression lente, et incluant une démarche instable, une dysarthrie, ainsi que des symptômes visuels tels qu'un nystagmus. D'autres traits sont rapportés dans plusieurs types, tels qu'un vertige important, une atrophie cérébelleuse, ou encore une faiblesse musculaire (37). La classification nécessite donc généralement une combinaison de plusieurs phénotypes précis (Tableau I), mais n'exclue pas la possibilité de pénétrance variable pour des mutations d'un même gène. En plus des huit formes mentionnées, l'ataxie épisodique avec épilepsie néonatale est parfois classifiée comme le 9^{e} type d'EA malgré les complications développementales atypiques à ce type de désordre (39). De façon similaire, un diagnostic du syndrome d'ataxie cérébelleuse, aréflexie, pieds creux, atrophie optique et perte auditive neurosensorielle (CAPOS) est souvent considéré dû à sa composante épisodique, mais la présentation symptomatique globale est très distincte d'une ataxie épisodique typique (40).

Tableau I. Présentation clinique des différents types d'ataxie épisodique (37)

DÉSORDRE	LOCUS ASSOCIÉ	SYMPTÔMES CARACTÉRISTIQUES
EA1	*KCNA1*	Myokimie; épisodes causés par des exercices; pas de vertige
EA2	*CACNA1A*	Nystagmus; épisodes causés par changement de posture; vertige
EA3	1q42	Vertige; acouphènes fréquents
EA4	-	Poursuite visuelle anormale; diplopie; malaises intenses
EA5	*CACNB4*	Symptômes précoces/adolescents; crises myocloniques
EA6	*SLC1A3*	Migraines causant photo/phonophobie; diplopie; Symptômes enfance
EA7	19q13	Symptômes précoces/adolescents; vertige; faiblesse musculaire
EA8	*UBR4*[2]	Faiblesse généralisée
EA9[1]	*SCN2A*	Épilepsie néonatale; autisme; hypo/dystonie
CAPOS	*ATP1A3*	Aréflexie; atrophie optique; perte auditive neurosensorielle

[1]Classification peut varier; [2]Candidat en cours de validation fonctionnelle

1.3.4.2 Connaissances génétiques

En regardant les gènes associés aux EA, il est intéressant de noter que la quasi-totalité est reliée aux échanges ioniques transmembranaires. Ce type de fonction corrèle avec l'aspect sporadique de la maladie, où les événements déclencheurs de crises ataxiques causent fort probablement un changement anormal des potentiels électriques membranaires (38). Évidemment, l'importance des ions dans la transmission de signaux neuronaux, via la polarisation et la dépolarisation des cellules pour la relâche de neurotransmetteurs comme le glutamate, est bien connue (41). L'ataxie épisodique étant une classe de maladie rare, sa fréquence est inférieure à 1/100 000, et la vaste majorité des cas répertoriés appartiennent aux sous-types 1 et 2 (37). La sous-unité Kv1.1 du canal potassique voltage-dépendant (*KCNA1*) est une protéine participant au transport transmembranaire d'ions K^+ dans le système nerveux central et les reins. Son rôle dans la régulation de l'excitation neuronale a été largement étudié, il n'est donc pas surprenant que sa perte de fonction engendre une communication inefficace entre les neurones hyperactifs (42). La majorité des mutations causant une EA1 sont faux-sens. Dans le cas de *CACNA1A*, le type de mutation pathogénique est plus variable, avec la plupart des EA2 causés par des changements du cadre de lecture ou des variants non-sens (43). Cependant, l'interprétation est généralement plus complexe, puisque l'altération du gène peut également engendrer une épilepsie juvénile, ou encore une SCA6. Au niveau fonctionnel, *CACNA1A* contribue à la dépolarisation des neurones excitables en permettant l'entrée cytoplasmique d'ions calcium (44). Participant au même complexe protéique, la sous-unité bêta 4 du canal calcique voltage-dépendant (*CACNB4*) cause plutôt l'EA4, du moins chez une famille canadienne portant un variant faux-sens (45). Similairement, des mutations pathogéniques ont été observées dans le gène du transporteur de glutamate *SLC1A3* chez quelques familles (46, 47). En plus d'une perturbation probable de la signalisation neuronale, il a été démontré que la protéine mutée pouvait engendrer des phénotypes via l'apoptose de cellules gliales. Dans le cas des gènes *SCN2A* et *ATP1A3*, leurs fonctions de transport d'ions ont un rôle à jouer dans le développement du cerveau, ce qui explique les phénotypes plus graves des maladies associées (39, 40). Finalement, le gène *UBR4* participe plutôt au renouvellement protéique et à l'organisation cytosquelettique via sa fonction de ligase à ubiquitine (48). La validation de la famille affectée est toujours en cours.

1.3.4.3 Variabilité

Un aspect récurrent entourant les différentes formes d'ataxie est l'hétérogénéité clinique plus que considérable qui peut être observée chez des patients porteurs de variants similaires dans un gène commun (38). En effet, même en examinant plusieurs patients d'une même famille, les différences dans la présentation phénotypique sont une quasi-certitude. Comme mentionné précédemment, cela représente un obstacle supplémentaire à la découverte de nouvelles causes génétiques. En effet, l'hypothèse d'une mutation commune peut être très logique, mais ne peut jamais complètement effacer la possibilité de patients atteints de pathologies divergentes d'un point de vue moléculaire. En revanche, un autre concept peut aider à l'interprétation de cette importante variabilité symptomatique : les modificateurs génétiques (49). Certaines altérations génomiques, bien qu'elles ne soient pas nécessairement pathogènes en eux-mêmes, ont la capacité de modifier les conséquences phénotypiques d'une mutation primaire. Il est intéressant de noter que bien que l'accent ait été mis sur les variants qui exacerbent la gravité de la maladie, appelée amplificateur, des variants de suppression ayant la capacité d'améliorer le pronostic des patients ont également été décrits (49). Ces derniers sont cependant légèrement plus difficiles à valider et ont un potentiel thérapeutique moindre, d'où la raison pour laquelle l'accent est généralement mis sur l'identification de mutations amplifiant la gravité de la maladie.

1.3.4.4 Traitements

Comme dans les autres maladies neurodégénératives, il n'existe malheureusement pas de thérapie efficace pour ralentir ou arrêter la progression clinique dans la plupart des cas. À l'exception des ataxies d'origine métabolique où des carences nutritionnelles sont en cause, ou des cas d'origine virale où le traitement de la source pathogénique externe est traitable, les seules médications disponibles pour la gestion des soins de patients permettent d'apaiser certains symptômes précis (38). Par exemple, certaines formes d'ataxies répondent positivement à la prise d'acétazolamide, un inhibiteur d'anhydrase carbonique causant une acidose métabolique et pouvant réduire la fréquence ainsi que l'intensité des symptômes ataxiques et épileptiques. Cependant, la plupart des patients semblent développer une résistance éventuelle au traitement, allant parfois jusqu'à l'induction d'hyperammoniémie suite à l'utilisation prolongée pour certaines formes d'ataxies comme l'EA2, où les symptômes sont subséquemment aggravés (50).

D’autres options thérapeutiques étant parfois utilisées incluent le riluzole, réduisant l’hyperexcitabilité neuronale via plusieurs mécanismes bloquant la transmission de signaux nerveux, des inhibiteurs glutaminergiques, ou encore des bloqueurs de récepteurs nicotiniques. Les thérapies de modulation neuronales, soient-elles sérotoninergiques, GABAergiques, ou encore cholinergiques, suggèrent certains bénéfices dans des cas précis, mais les résultats ne sont pas concluants (51). La gestion des soins de patients ataxiques inclut plus fréquemment des sessions de thérapies physiques et linguistiques lorsqu’applicable. En effet, malgré l’absence de traitement efficace, plusieurs études soulignent le potentiel de la neuro-réhabilitation dans le ralentissement de la progression pathogénique, particulièrement chez les patients présentant des symptômes précoces ou légers (51, 52). Il est même possible que la plasticité neuronale naturelle soit en mesure d’améliorer les capacités physiques dans certains cas où les activités de réhabilitations sont un succès.

1.4 Maladies neurologiques connexes

Comme mentionné plus tôt, en plus des maladies neurologiques avec un phénotype d’ataxie en premier plan, plusieurs désordres génétiques ont une présentation clinique largement différente, mais incluent des troubles d’incoordination secondaires. Il est donc utile de connaître les gènes causatifs de ces maladies afin de pouvoir considérer leur implication dans des cas complexes d’ataxies épisodiques avec symptômes hétérogènes.

1.4.1 Paraplégies spastiques

Les paraplégies spastiques héréditaires (HSP) sont un groupe de maladies neurologiques affectant également la démarche du patient. Cependant, comme son nom l’indique, la caractéristique principale des HSP est une importante spasticité des membres inférieurs (53). La raideur n’est toutefois pas toujours restreinte aux jambes, et les autres symptômes typiques mais non obligatoires du désordre incluent une faiblesse musculaire, une hyperréflexie, ainsi qu’une diminution de la perception sensorielle des membres inférieurs. Il existe plus de 80 types de paraplégies spastiques, il n’est donc pas étonnant que l’âge d’apparition symptomatique varie de présentation néonatale à tardive. Cependant, il est intéressant de noter que dans la plupart des cas, l’apparition de symptômes à un très jeune âge est associée à des phénotypes non progressifs,

alors qu'une progression lente et constante est plutôt associée aux autres présentations (53). Les deux classes majeures de paraplégies spastiques sont définies comme HSP pure, où les phénotypes se limitent principalement à la présentation typique au niveau des membres inférieurs, et HSP complexe, où les déficiences impliquent plusieurs systèmes. On y retrouve entre autres des symptômes d'ataxie, d'épilepsie, de déficience intellectuelle, de démence, d'atrophie musculaire, ainsi que des neuropathies périphériques. Encore une fois, il n'est pas surprenant d'apprendre qu'un gène, tel que *SPG7*, peut parallèlement être en cause pour les deux types de HSP (54). Malgré le fait que près de 80% des HSP soient autosomiques dominantes, de nombreuses formes AR et liées au chromosome X ont été décrites. Parmi les nombreux gènes causatifs identifiés, un peu moins d'une vingtaine sont reconnus pour induire des phénotypes ataxiques, et sont pertinents à l'analyse génétique de patients présentant des troubles cérébelleux tels que l'EA (53).

1.4.2 Désordres mitochondriaux

Un autre type de maladie complexe qui présente fréquemment des symptômes d'ataxie secondaire sont les troubles mitochondriaux. D'un point de vue moléculaire, ce type de maladie résulte de mutations dans des gènes importants pour le bon fonctionnement de la chaîne respiratoire mitochondriale (55). Comme les neurones sont parmi les cellules les plus énergivores du corps humain, il est logique que des phénotypes neurologiques soient souvent observés dans ce type de pathologie (56). Les autres caractéristiques cliniques courantes comprennent des anomalies oculaires, des myopathies, des neuropathies et le diabète de type I. Alors que l'ADN mitochondrial (ADNmt) code la plupart des gènes importants pour les fonctions des mitochondries, le génome nucléaire (ADNn) contient également de nombreux gènes qui régulent l'expression de l'ADN mitochondrial, participent à la synthèse des protéines ou jouent un rôle direct dans la chaîne respiratoire (55). Sans entrer en détail dans les conditions induites par des mutations dans ces deux types d'ADN, de multiples troubles mitochondriaux ont des gènes causals étant pertinents pour les patients ataxiques, notamment dans les syndromes de Kearns-Sayre et de Leigh, dans la faiblesse neurogène avec ataxie et rétinite pigmentaire (NARP), ainsi que dans l'épilepsie myoclonique à fibres rouges déchiquetées (MERRF) (57). Dans l'ensemble, cela représente 40 à 50 gènes supplémentaires à prendre en compte chez des patients présentant

des phénotypes ataxiques. Cependant, bien qu'il soit important de les considérer lors de l'analyse génétique, l'absence d'une caractéristique distinctive d'une maladie donnée peut permettre d'éliminer l'implication d'un gène précis dans la pathologie du patient. La clé de l'interprétation des variants demeure la revue de la littérature disponible.

1.4.3 Autres

Il existe de nombreux autres troubles neurologiques rares provoquant des symptômes d'ataxie secondaire qui ne sont pas toujours pertinents pour un cas spécifique, mais qui peuvent être pris en compte lorsqu'un patient présente des symptômes cérébelleux. Les deux formes de maladies de Niemann-Pick type C, qui sont classées comme des troubles lysosomaux affectant le métabolisme des lipoprotéines et du cholestérol, induisent généralement des phénotypes d'ataxie (58). En effet, alors que les anomalies proviennent d'un dysfonctionnement du foie et de la rate, l'accumulation de sphingomyéline qui en résulte provoque de multiples symptômes liés au cerveau, y compris ceux généralement associés à une dégénérescence du cervelet. Similairement, le spectre de Zellweger définit des troubles résultant de peroxysomes dysfonctionnels, un organite responsable de la dégradation des acides et d'autres composés toxiques (59). Plusieurs gènes peuvent provoquer ce type de maladie, tels que *PHYH*, *PEX2* ou *PEX7*, où des mutations pathogéniques conduisent à des phénotypes neurologiques sévères (60). Le point important à retenir est que la littérature pertinente à cette recherche dépasse les formes typiques d'ataxie, et peut considérablement améliorer les capacités d'interprétation de variants. Par conséquent, il est pertinent de rassembler l'information sur les gènes liés à l'ataxie de près ou de loin (Tableau II) afin d'augmenter l'efficacité d'analyse.

Tableau II. Gènes ayant la capacité de causer des phénotypes d'ataxie (16, 27, 53, 55)

Ataxie primaire AD (33 gènes[1])	*AFG3L2; ATN1; ATXN1; ATXN2; ATXN3; ATXN7; ATXN8OS; ATXN10; BEAN1; CACNA1A; CACNA1G; CACNB4; CCDC88C; EEF2; ELOVL4; FGF14; ITPR1; KCNA1; KCNC3; KCND3; NOP56; PDYN; PPP2R2B; PRKCG; RAB11B; SCA21; SCA25; SLC1A3; SPTBN2; TBP; TGM6; TTBK2; VAMP1*
Ataxie primaire AR (70 gènes[1])	*ABHD12; AFG3L2; ANO10; APTX; ATCAY; ATM; CA8; CAPN1; CLCN2; COA7; COQ8A; COX20; CWF19L1; CYP27A1; CYP7B1; DARS2; DNAJC19; DNAJC3; FLVCR1; FXN; GBA2; GDAP2; GJC2; GOSR2; GRID2; GRM1; GRN; ITPR1; KCNJ10; KIF1C; L2HGDH; LAMA1; MRE11A; MTPAP; PCNA; PEX10; PIK3R5; PMPCA; PNKP; PNPLA6; POLG; POLR3A; POLR3B; RFC1; RNF216; RUBCN; SACS; SCYL1; SETX; SIL1; SLC9A1; SNX14; SPG7; SPTBN2; STUB1; SYNE1; SYT14; TDP1; TDP2; THG1L; TPP1; TTPA; TWNK; UBA5; UCHL1; VLDLR; VPS13D; VWA3B; WDR81; XRCC1*
Ataxie secondaire (66 gènes)	*AHI1; ALDH5A1; ALG6; ARL13B; ATP7B; ATP8A2; B4GALNT1; BTD; CC2D2A; CEP290; CLN5; CLN6; CP; CSTB; CTSD; EIF2B1; EIF2B2; EIF2B3; EIF2B4; EIF2B5; EPM2A; ERCC4; FA2H; FOLR1; FRMD4A; GAN1; GLB1; HEXA; HEXB; HSD17B4; KCTD7; MAN2B1; MLC1; MSTO1; MTTP; NEU1; NHLRC1; NKX6-2; NPC1; NPC2; OFD1; OPA1; PEX2; PEX7; PHYH; PLA2G6; PMM2; PRRT2; PTRH2; RPGRIP1L; SEMA6B; SEPSECS; SLC2A1; SLC5A2; SLC6A19; SLC17A5; SLC25A46; SPG11; SRD5A3; TMEM67; TMEM231; TTC19; WDR45; WDR73; WFS1; WWOX*

[1] Si un gène est reconnu comme pouvant causer les deux types, il n'est identifié qu'une fois.

1.5 Séquençage de Nouvelle Génération

Les approches de séquençage de l'ADN ont largement évolué au cours des dernières décennies, tout comme le nombre d'applications du séquençage à haut débit en sciences fondamentales et cliniques. Alors que le séquençage de nouvelle génération est le terme le plus couramment utilisé pour les technologies qui ont succédé au séquençage original par Sanger, il regroupe en fait un large éventail de méthodes différentes qui se sont elles-mêmes améliorées au fil des ans. À l'origine, NGS désigne les séquençages par hybridation et par synthèse, ce dernier étant la technique la plus connue due aux nombreuses itérations du séquençage Illumina (61). Centrés sur le principe de l'amplification en pont sur matrice, qui contourne les besoins traditionnels d'électrophorèse et de clonage bactérien, de nombreux protocoles ont été développés pour évaluer les différentes caractéristiques moléculaires des multiples types de séquences d'acides nucléiques présentes dans les cellules. Parmi les méthodes les plus populaires figurent le

séquençage du génome entier, de l'exome, ciblé pour définir des séquences d'ADN, le séquençage transcriptomique renseignant sur les ARN messagers (ARNm), et le séquençage du méthylome donnant un aperçu de la régulation épigénétique. Récemment, de nouvelles technologies permettant la lecture de plus longs fragments nucléotidiques ont été décrites comme la troisième génération de séquençage, parallèlement au séquençage à cellule unique. Des entreprises telles qu'Oxford Nanopore et PacBio ont développé différentes approches de séquençage à longues lectures (LRS) ayant un immense potentiel, autant pour des fins de recherche que pour la médecine personnalisée (62). Chaque méthode a ses avantages et ses limitations, il est donc important de définir les objectifs d'un projet pour sélectionner les technologies appropriées.

1.5.1 Séquençage génomique

Parmi les techniques de séquençage de l'ADN, le séquençage du génome entier (WGS) est l'approche qui génère le plus d'informations, car elle capture également la majorité des régions non codantes du génome humain (10). Historiquement, la recherche de mutations causales s'est concentrée sur les régions codantes, car la variation directe de la séquence d'un gène est le plus susceptible d'affecter sa fonction moléculaire. Néanmoins, bien que l'interprétation de ces variants ne soit toujours pas optimale, les méthodes d'identification actuelles axées sur les régions codantes donnent généralement un taux de réussite diagnostique d'environ 40%, suggérant une implication considérable du 98% du génome restant (63). En fait, alors qu'il est appelé ADN « non codant », il est supposé que jusqu'à 82% de celui-ci pourrait être fonctionnellement pertinent, avec des éléments régulateurs composant une grande partie de ces régions (64). Cela rend le WGS très attirant, car une vue d'ensemble complète du génome humain donne le portrait le plus fidèle des causes possibles de la pathologie d'un patient. Dans cette méthode, l'ADN nucléaire extrait est généralement uniquement cisaillé de manière non spécifique avant que les adaptateurs soient ajoutés aux fragments pour permettre le séquençage subséquent (Figure 2). L'absence d'une étape de réaction en chaîne par polymérase (PCR) présente de multiples avantages, dont une plus grande fidélité de séquence, et une meilleure répartition de la couverture du génome. En effet, l'ADN humain contient de nombreux motifs provoquant des biais d'amplification, ce qui conduit généralement à une surreprésentation des séquences portant ces motifs dans les données de sorties (65). Bien que les avantages du WGS

soient évidents, l'approche ne vient pas sans un ensemble considérable de limitations. En effet, comme l'interprétation des changements de nucléotides est encore une tâche compliquée dans les séquences codantes, une couche de complexité supplémentaire s'ajoute pour des régions qui ont des rôles indéfinis dans la régulation des gènes, et où les règles définissant l'architecture fonctionnelle sont encore profondément incomprises (63). Alors que les experts dans le domaine de la bio-informatique développent rapidement des outils d'interprétation plus efficaces et pour davantage d'applications, exploitant fréquemment les progrès récents de l'apprentissage par machine où l'intelligence artificielle est très prometteuse (66), il faudra un certain temps aux algorithmes pour prédire correctement les effets pathogéniques des variants non codants (67). En ce qui concerne les pipelines d'analyse présentement disponibles, seuls les variants codants ont tendance à passer les filtres de pertinence. Cela est dû à la programmation des outils, qui sont principalement construits à partir de la littérature publiée sur les effets fonctionnels résultant de mutations variées, mais où la majorité des données d'entraînement proviennent du génome codant (68, 69). Cela étant dit, la prédiction de la pathogénicité à partir d'observations extrapolées n'est pas la seule approche disponible permettant d'identifier des changements fonctionnels pertinents qui découlent de mutations non codantes. Il a été récemment proposé que la combinaison de plusieurs technologies à haut débit, ou « multiomique », pourrait conduire à une évolution rapide de notre compréhension des éléments génomiques régulateurs et de la façon dont certaines mutations sont capables d'induire des phénotypes en altérant seulement l'expression génique (70, 71). La transcriptomique désigne l'étude de l'expression des gènes, et représente donc un excellent candidat pour l'obtention d'informations fonctionnelles sur l'état moléculaire d'un patient dont la séquence génomique complète est disponible suite au WGS.

1.5.2 Séquençage transcriptomique

Le séquençage d'ARN (RNA-seq) à lectures courtes d'Illumina est une technique populaire utilisée pour évaluer quantitativement la présence de diverses molécules d'ARN dans une cellule (72). L'ARN total est obtenu à partir d'une simple élimination du contenu en ADN, soit par DNase ou bien par chromatographie sur colonne. Cela offre évidemment de l'information sur les ARNm, mais aussi sur d'autres ARN non codants comme les micro-ARN (miARN), les ARN nucléolaires courts (snARN), les ARN de transfert (ARNt), les ARN ribosomiques (ARNr) et les ARN longs non

codants (lncARN) (73). Malgré des modifications post-transcriptionnelles telles que l'ajout d'une coiffe en 5' et la polyadénylation en 3', cette dernière étant parfois exploitée pour l'enrichissement d'ARNm codant par oligo dT, les molécules d'ARN sont très instables (74). Par conséquent, la synthèse d'ADN complémentaire (ADNc) est une étape presque essentielle de la préparation d'une librairie de séquençage, qui permet en plus une amplification ARN présents pour augmenter la couverture de séquençage. Il est important de noter que cette étape peut introduire des biais et des artefacts indésirables qui peuvent affecter négativement les annotations ou la quantification des transcrits (73). Cependant, les méthodes alternatives telles que le séquençage direct de l'ARN via Nanopore comportent également leurs limites, notamment des coûts accrus pour une couverture plus faible (61). Encore une fois, de nombreuses itérations de la méthode, telles que le séquençage de miARN ou encore unicellulaire, ont été développées pour leurs avantages spécifiques dans certaines situations (73). Selon les objectifs d'une étude, une décision éclairée doit être prise pour sélectionner l'approche appropriée. Un autre facteur à considérer est le tissu à partir duquel les macromolécules sont obtenues. En effet, contrairement au séquençage génomique, l'ARN doit être exprimé pour sa détection par RNA-seq. L'expression des gènes est un processus fortement régulé et de nombreux transcrits sont produits de manière tissu-dépendante (75). Alors que le tissu d'intérêt est généralement préféré pour l'obtention d'un juste portrait des mécanismes moléculaires en jeu dans la pathologie d'un patient, la disponibilité du matériel cellulaire peut parfois nécessiter un prélèvement très invasif. Dans le cas des troubles neurologiques, les cellules cérébrales sont simplement inaccessibles; il est possible que les gènes pertinents ne soient pas exprimés dans le prélèvement. Une autre considération importante est la couverture de séquençage souhaitée, puisque la quantification de l'abondance de l'ARNm et la détection d'événements tels que l'épissage alternatif nécessitent de nombreuses lectures par transcrits afin de pouvoir discerner les changements réels de simples artefacts (72).

L'interprétation complète des données RNA-seq nécessite des connaissances sur le traitement de l'ARNm : la transcription de la séquence d'ADN d'un gène entraîne la formation d'un pré-ARNm, qui contient encore des séquences introniques devant être éliminées par les spliceosomes (76). Ces grands complexes ribonucléoprotéiques reconnaissent des éléments de séquence spécifiques présents dans les introns, y compris des motifs proches des jonctions d'épissages en 5' et en 3'.

Ces motifs sont généralement très conservés, ce qui se traduit sans surprise par un épissage anormal lorsque des mutations surviennent dans ces régions (77). L'interprétation des variants plus profondément introniques est moins évidente, mais l'altération de point de branchement au spliceosome peut par exemple également engendrer la formation de nouveaux transcrits. Alors que l'épissage alternatif est un processus important ayant grandement participé à la génération de la diversité fonctionnelle génétique, ces événements canoniques sont régulés avec précision (72). Certains isoformes d'ARNm d'un même gène peuvent avoir différentes fonctions moléculaires, et l'expression de certains transcrits peut être spécifique à un ou plusieurs tissus. Ce type de phénomène donne un aperçu de la précision de régulation nécessaire à l'épissage, et des répercussions sur diverses voies cellulaires que peut engendrer un simple changement dans l'équilibre des isoformes, incluant le développement de pathologies du cervelet (78).

1.5.3 Séquençage à longues lectures

L'une des limitations du RNA-seq décrit est la longueur des lectures, qui couvrent rarement plusieurs jonctions d'exons. Cela signifie que les transcrits sont assemblés à partir de prédictions générées suite à l'empilement de lectures présentant des similitudes dans leur séquence et en comparaison avec des annotations de référence contenant habituellement les transcrits canoniques ou plus fréquents. Bien que les outils d'alignement aient fait des progrès impressionnants en termes de précision au courant des dernières années, plusieurs articles suggèrent maintenant que les annotations disponibles ne sont pas aussi précises que prévu, du moins de manière tissu-spécifique, et la quantification d'isoformes est toujours très difficile (79). Par conséquent, le LRS permettant des lectures plus de cent fois plus grandes est susceptible de conduire à des améliorations considérables des annotations des transcrits de nombreux gènes grâce à des assemblages *de novo*, ce qui devrait à son tour être bénéfique pour le séquençage d'ARN en général (62). L'originalité de la méthode provient de l'absence d'une baisse de qualité progressive lors de l'étape de séquençage. En effet, avec les méthodes traditionnelles, les signaux luminescents proviennent de l'amplification cyclique de fragments d'ADN, où les brins amplifiés deviennent de plus en plus asynchrones à chaque cycle jusqu'à ce que les différents nucléotides soient indiscernables (61). Avec les technologies LRS développées par PacBio et Oxford Nanopore, la lecture d'un fragment d'ADN ne nécessite pas d'amplification du signal par amas, éliminant

ainsi le problème de déphasage. Par exemple, le séquençage Nanopore exploite des protéines transmembranaires pour créer des pores à travers une membrane lipidique, permettant le passage de fragments d'ADN natifs (80). Pour augmenter le taux de séquençage, les complexes poreux comprennent également des protéines motrices ainsi que des hélicases pour dérouler l'ADN double brin. Un capteur d'état solide est ensuite capable d'interpréter le signal nucléotidique traversant les pores. Bien sûr, d'autres problèmes peuvent survenir lors du passage de molécules uniques, tels qu'un manque de stabilité des grands fragments d'ADN ou le blocage des pores avec l'approche d'Oxford, mais des lectures de plusieurs mégabases ont déjà été accomplies (61). Le compromis vient plutôt de la précision de l'appel de base, où les signaux générés correspondant aux nucléotides ne sont pas définis aussi clairement avec les outils LRS qu'avec ceux du séquençage traditionnel à lectures courtes (81). Par conséquent, une meilleure couverture est nécessaire pour une détermination fiable des séquences, ce qui peut augmenter le coût expérimental. D'autre part, les séquenceurs tels que le MinION sont facilement portables en raison de leur petite taille, ce qui les rend très intéressants dans les situations où l'accès à d'autres machines ou bien l'espace est limité (61). De plus, le LRS est déjà très polyvalent dans ses applications possibles, avec des protocoles disponibles pour le multiplexage d'échantillons, le séquençage du méthylome et l'accessibilité de régions qui sont généralement difficiles à cibler ou à amplifier (82). Parmi ceux-ci se trouvent les régions de répétitions en tandem court (STR), où de nombreux exemples d'expansion nucléotidique affectant l'expression de gènes à proximité, induisant subséquemment des maladies chez l'humain, ont été rapportés (31). Ces régions peuvent souvent s'étendre sur plusieurs milliers de nucléotides, ce qui rend impossible pour les lectures courtes de définir la séquence complète et le nombre de répétitions en cause. Traditionnellement, cette information doit être déterminée par buvardage de Southern, qui est une méthode laborieuse et coûteuse. Le LRS s'avère donc être une approche efficace pour déterminer la nature et la longueur des séquences d'expansions répétées pathologiques. La découverte de ce type d'applications utiles offertes par le séquençage de longues lectures a suscité un énorme optimisme dans la communauté des technologies omiques, conduisant à un développement rapide d'outils bio-informatique pour l'analyse des données générées. Bien qu'encore une phase de développement précoce, le LRS a un immense potentiel dans de

nombreux domaines de la recherche fondamentale et clinique (82). Entre autres, la technologie risque de s'avérer très utile pour définir plus précisément le transcriptome de nombreux gènes pour lesquels les isoformes d'épissage ne sont pas encore tous décrits à ce jour (62).

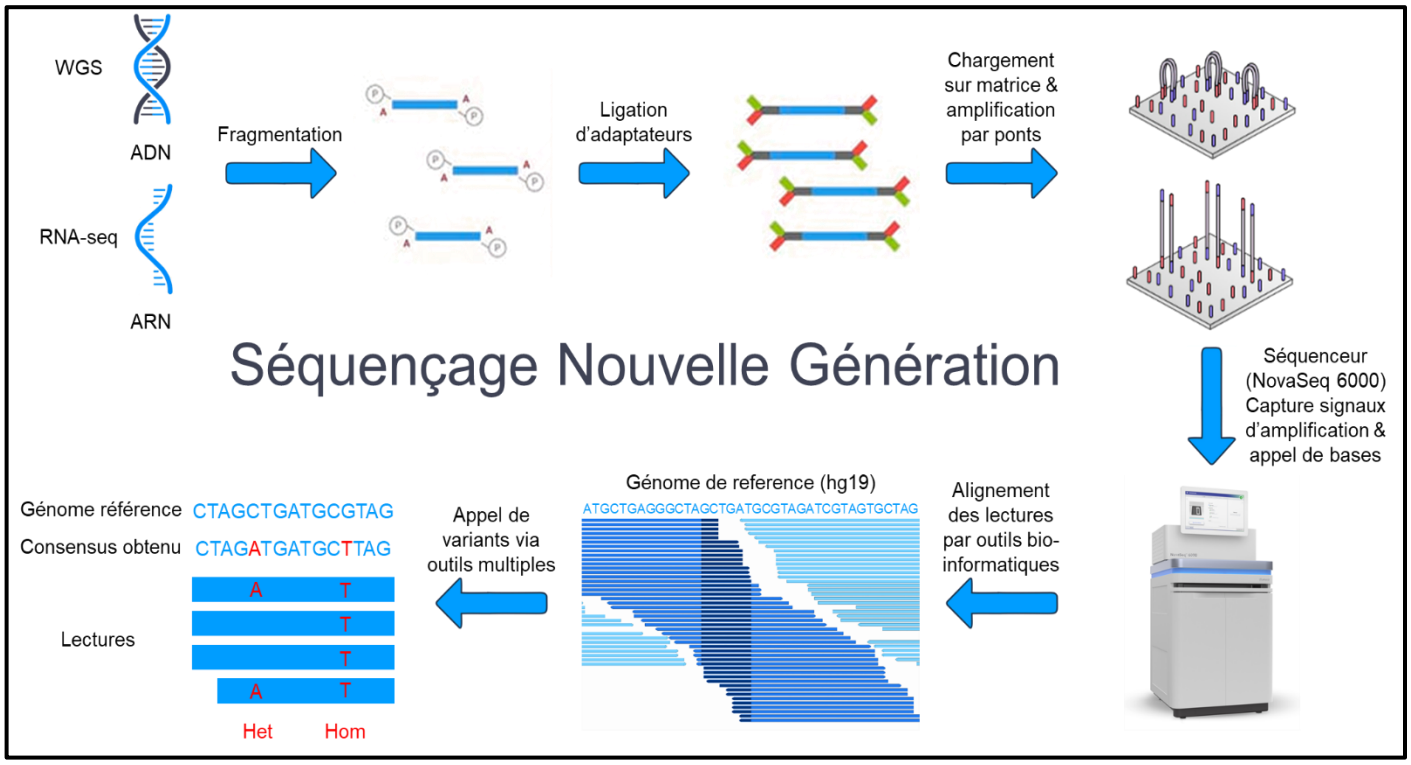

Figure 2. Représentation visuelle du processus de séquençage nouvelle génération. Schéma simplifié des étapes nécessaires à l'obtention de données de séquençage pour analyse génétique. Certaines étapes tels que les contrôles qualités ne sont pas représentées.

1.5.4 Outils bio-informatiques

Avec l'avènement des technologies à haut débit, la demande d'outils bio-informatiques efficaces permettant une analyse de données simple et rapide a augmenté de façon exponentielle. Puisqu'il permet de construire un portrait représentatif de l'état des séquences d'ADN ou d'ARN de cellules humaines, le NGS est l'une des technologies générant la plus grande quantité de données, atteignant parfois plusieurs centaines de giga-octets pour une expérience. Non seulement est-il important de traduire les signaux captés par le séquenceur en informations génomiques interprétables, mais il est essentiel de trouver des moyens d'extraire spécifiquement les parties d'information pertinente à une question de recherche afin de limiter de façon réaliste

les ressources requises pour l'investigation. Quelle que soit la technologie de séquençage, les pipelines d'analyses se chevauchent partiellement pour une branche commune où les fichiers initiaux (FASTQ) contenant les lectures sont alignés sur un génome de référence et annotés avec des caractéristiques génétiques connues. Les outils bio-informatiques notables dans l'accomplissement de ce traitement comprennent un aligneur, des filtres de lectures non conformes, des outils d'évaluation de qualité, et des bases de données d'annotations (83). En fonction de l'objectif d'une expérience, les outils subséquents sont choisis afin de prédire divers événements et visent à identifier les caractéristiques atypiques d'un échantillon pouvant être pertinentes pour l'utilisateur. Parmi ceux-ci se trouvent l'appel de mutation pour les différents types de variants tels que les SNV, les SV, les AS ou bien les expansions nucléotidiques, ainsi que la quantification des gènes du RNA-seq afin d'effectuer l'analyse d'expression différentielle (84).

1.5.4.1 Prétraitement des fichiers LRS de départ

Comme mentionné précédemment, l'appel de base LRS n'a pas encore atteint une efficacité optimale, et n'est donc pas toujours effectué en temps réel contrairement au séquençage Illumina. Bien que l'amélioration des algorithmes actuels est à prévoir, Guppy s'est établi comme outil puissant pour le prétraitement des données LRS (85). Plusieurs protocoles sont disponibles pour permettre de choisir entre un appel de base en temps réel et un appel de haute précision des données générées par le séquenceur (FAST5). Ce dernier offre une précision nucléotidique d'environ 88 à 90 % sur les lectures individuelles, ce qui est convenable si l'on considère qu'une bonne couverture permet d'obtenir une identité de consensus atteignant plus de 99,40 % (85). Bien sûr, une meilleure précision s'accompagne d'un léger compromis sur la vitesse d'exécution, mais les performances demeurent comparables à celles des outils homologues. De plus, Guppy offre des fonctions de prétraitement supplémentaires telles que le démultiplexage, qui nécessite autrement l'installation d'outils complémentaires tels que DeepBinner (86). Cela permet à l'utilisateur de charger plusieurs échantillons sur une même cartouche de séquençage, à condition que les lectures aient été marquées avec des codes-barres nucléotidiques. Par conséquent, il est actuellement considéré comme faisant partie intégrante du traitement des données LRS standard, conduisant à la génération de fichiers FASTQ appropriés (62).

1.5.4.2 Alignement des lectures sur un génome de référence

L'étape d'alignement est cruciale pour l'analyse des données NGS, puisqu'elle transforme des millions de lectures aléatoires en séquences consensus représentatives de l'identité moléculaire de l'échantillon. Par conséquent, les outils d'alignement ont été largement optimisés pour assurer une combinaison optimale de vitesse de traitement et de précision pour tout type de données. En considérant que la source des fragments séquencés ainsi que la longueur des lectures peuvent affecter la logique conduisant au meilleur alignement sur un génome de référence, les outils standards diffèrent en fonction des technologies de séquençage. Pour le WGS, l'aligneur Burrow-Wheeler (BWA) est la méthode de choix, la nouvelle itération BWA-Mem étant généralement plus performante et rapide pour la majorité des données génomiques (87). Le développement d'une version mise à jour (BWA-Mem2) met en évidence l'intérêt d'améliorer les performances d'un outil jusqu'à l'atteinte de la quasi-perfection. Pour les données transcriptomiques, l'aligneur standard n'est pas défini aussi clairement, avec certains outils présentant une meilleure précision et sensibilité d'alignement tandis que d'autres offrent des avantages évidents en termes de temps d'exécution. À cet effet, il existe des pseudo-aligneurs comme Kalisto et Salmon qui (88, 89), au lieu d'aligner chaque nucléotide d'une lecture au génome de référence, fragmentent les données en k-mers afin d'effectuer moins de comparaisons statistiques, permettant une quantification plus rapide des différents transcrits (90). Bien qu'ils ne soient pas les plus précis, deux des méthodes d'alignement les plus populaires sont STAR et HISAT2, ce qui est probablement dû au fait qu'elles conservent de très bonnes performances tout en étant les aligneurs les plus rapides (91, 92). De plus, ils sont assez faciles à utiliser, ne nécessitent pas d'annotations pour traiter les lectures et sont évidemment sensibles à l'épissage (splice-aware) (93). En effet, la principale différence avec les aligneurs d'ADN est que les algorithmes sont programmés pour considérer que de nombreuses lectures doivent être divisées pour tenir compte du retrait des introns dans l'ARNm. Une autre considération importante affectant le RNA-seq est l'alignement de lectures sur plusieurs sites, également connue sous le nom de cartographie ambiguë. Alors que l'effet de celle-ci sur l'alignement génomique est négligeable, elle affecte la précision de quantification des transcrits, information utile en termes d'évaluation fonctionnelle de voies moléculaires ou de gènes spécifiques. Par conséquent, le compromis entre sensibilité et spécificité peut déterminer

l’algorithme d’alignement le plus approprié à un contexte spécifique. Alors que STAR présente une sensibilité supérieure lorsqu'il est correctement configuré, ce qui peut être mieux pour un appel de variant précis, HISAT2 offre apparemment une meilleure spécificité d'alignement pouvant bénéficier la quantification ultérieure des gènes (93, 94). Naturellement, il est possible d'utiliser deux algorithmes en parallèle pour assurer l'exactitude des résultats, au détriment du temps de traitement. Pour les données LRS, minimap2 s’est récemment établi comme un aligneur puissant pour les longues lectures, pouvant être finement ajusté en fonction de la technologie utilisée, et offrant de bonnes performances à la fois pour Oxford Nanopore et PacBio (81). Développé par la même équipe qui a conçu BWA, l'outil disposait déjà d'une bonne base, avec de multiples fonctionnalités importées de BWA-Mem. L’ajout de fonctionnalités supplémentaires telles qu'un algorithme de chaînage plus rapide et de nouveaux modèles pour l’évaluation des écarts d’alignement permettent donc à cet outil de se séparer du lot (95).

1.5.4.3 Contrôle de la qualité des données

Comme les informations extraites des données NGS influencent généralement soit une étude fondamentale, soit les patients directement via la recherche clinique et le diagnostic, l'assurance qualité est essentielle pour éviter des conclusions trompeuses (83). Il a déjà été établi que plusieurs points de contrôle qualité évitent des problèmes majeurs lors de la préparation et du séquençage de la librairie, mais d'autres anomalies peuvent être détectées durant le traitement des données avec des outils bio-informatiques. Par exemple, l'observation d'une grande proportion de scores phred faibles (score Q, niveau de confiance de l'appel de base) dans les données FASTQ peut généralement être attribuée à des erreurs du séquenceur, ou bien des problèmes survenus lors du chargement de l'échantillon. Un nombre anormalement élevé de lectures non alignées dans les fichiers BAM peut suggérer une contamination de la librairie par diverses sources d'acides nucléiques, étant fortement susceptible d'interférer avec l'analyse ultérieure (96). Ces informations, ainsi que d’autres détails tels que les effets de lots différentiels, sont obtenues à partir d'outils de qualité tels que MultiQC, qui regroupe plusieurs analyses dans un seul rapport interactif avec des graphiques faciles à interpréter (97). Cependant, comme il a été développé pour des données de lecture courtes, d'autres outils sont mieux adaptés au LRS. PycoQC est particulièrement efficace pour effectuer l'évaluation qualitative, puisqu’il offre une

interactivité bonifiée avec des graphiques fluides, et a été conçu spécifiquement pour les données d'Oxford Nanopore (98). Une autre information pertinente étant détectée par les outils de contrôle qualité est la présence d'adaptateurs de séquençage dans les lectures. Bien que le clivage d'adaptateurs soit partiellement intégré à la plupart des aligneurs modernes via des fonctions d'écrêtage, les résultats ne sont pas parfaits, ce qui peut entraîner à la fois un alignement erroné de séquences d'adaptation et la perte d'alignement pour des lectures ayant autrement une bonne qualité. Dans ce cas, un rognage d'adaptateur ciblé et précis peut être effectué avec des outils tels que Trimmomatic, qui est très populaire pour le retrait des adaptateurs Illumina (99). En fait, à moins que la conservation des adaptateurs ne soit requise dans un but précis, leur retrait avant l'alignement est recommandé. Durant ainsi qu'à la fin de l'étape d'alignement, les lectures sont étiquetées avec diverses métriques, utilisées pour capter et retirer les données de mauvaise qualité. Bien que certains de ces filtres soient inclus avec les aligneurs, les guides de meilleures pratiques recommandent généralement une notation additionnelle des lectures pour assurer une qualité optimale pour l'analyse ultérieure (83). Un bon exemple est le recalibrage du score de qualité de base, qui nécessite l'évaluation de l'ensemble de données post-alignement. L'objectif est de détecter les potentielles erreurs d'appel de base systématiques provenant du séquenceur et d'ajuster les scores phred en conséquence. Bien que l'impact soit négligeable dans la plupart des cas, l'exécution de cette étape supplémentaire avant l'appel de variant peut particulièrement améliorer la spécificité de la détection des SNV, bien que cela puisse réduire la sensibilité pour les régions à faible couverture (100).

1.5.4.4 Génération du fichier d'appel de variant avec annotations

Pour tout NGS à lecture courte, la boîte à outils d'analyse du génome (GATK) est la norme incontestée pour l'appel de variants, du moins pour la détection de SNV et de petites insertions/suppressions (indels) sur l'ensemble du génome humain (101). Avec une précision supérieure à 99%, il n'est généralement pas nécessaire de combiner des outils pour éviter les faux positifs. En effet, la plupart de ces artefacts, détectés comme des mutations par des outils comme HaplotypeCaller, sont supprimés lors du processus de filtrage. Cependant, la vérification d'un candidat par examen visuel avec des outils tels que le visualiseur génomique interactif (IGV) doit toujours être effectuée, car il ne nécessite que peu de temps et le filtrage automatique

n'attrapera pas tous les faux positifs (102). Parmi ceux-ci figurent les appels provenant de lectures de faible qualité, les artefacts de fin des lectures, les désalignements près des indels, les artefacts dus aux biais de brins et les désalignements près des régions de faible complexité (83). La vérification via séquençage de Sanger n'est plus une exigence absolue, puisque des études ont montré que la précision du GATK est supérieure à 99,96 %, mais reste standard lorsque les ressources sont disponibles pour effectuer cette étape supplémentaire (103). GATK offre un ensemble d'outils très polyvalent, avec une option très intéressante pour faire un appel joint lorsque les données d'un trio familial sont disponibles. Étant donné le fait que les mutations héritées peuvent être définies à partir des données parentales, l'identification des variants *de novo* dans le proband est simplifiée (101). Bien que ce type de mutation soit beaucoup plus rare que les variants héréditaires, ils représentent une proportion substantielle des diagnostics obtenus à partir des tests cliniques et doivent toujours être pris en compte lors de l'analyse (104). Suite à l'appel d'haplotype de GATK, le fichier de format d'appel de variant (VCF) obtenu est l'entrée standard pour ajouter des annotations concernant les caractéristiques des gènes associés aux variants. ANNOVAR est l'outil d'annotation le plus couramment utilisé, puisqu'il est facile à utiliser et récupère rapidement les informations de diverses bases de données telles que gnomAD pour la fréquence des allèles, dbSNP afin de savoir si un changement de nucléotide a précédemment rapporté dans d'autres tests cliniques, ou OMIM et ClinVar pour l'association de maladies connues (43, 105-108). D'autres annotations utiles incluent le type de mutation, l'effet attendu sur la protéine ou l'épissage, les fonctions connues du gène et les multiples scores d'importance des algorithmes qui prédisent le potentiel pathogénique des variants, tels que CADD, Polyphen2 ou SIFT (109-111). Ces derniers considèrent parfois l'aspect phylogénétique des mutations, où la conservation de nucléotide est fréquemment un signe de correspondance fonctionnelle. En plus de permettre l'interprétation de la pertinence d'une mutation pour un échantillon, les annotations fonctionnelles permettent un filtrage rigoureux des variants rares par la fréquence allélique et de meilleurs candidats par les scores de pathogénicité. En filtrant pour ces variants rares et majoritairement codants, le rapport obtenu à partir d'un VCF annoté extrait, à partir de plusieurs millions de résultats initiaux, autour de 100 mutations candidates adaptées à l'examen approfondi de leur association potentielle avec les phénotypes du patient.

1.5.4.5 Outils de prédictions supplémentaires

Comme mentionné précédemment, les SNV et les indels ne sont pas les seuls types d'altérations génomiques capables d'induire des changements pathogéniques dans les cellules humaines. Par conséquent, il est important que le pipeline d'analyse inclue des outils supplémentaires optimisés pour l'identification des autres types de mutations spécifiques. Le WGS à bouts appariés offre les meilleures chances d'identifier un SV, même si la précision n'est toujours pas excellente en raison de la nature des variants structurels. En effet, le fait que les lectures NGS soient plus courtes qu'un SV moyen, combiné à une localisation fréquente près des régions de faible complexité telles que les STR, rend leur détection difficile (112). Néanmoins, de nombreux outils ont été développés dans cette optique, DELLY et Lumpy étant deux des méthodes les plus populaires, exploitant un modèle de lectures-fractionnées (113, 114). L'autre type d'algorithmes utilisé pour la détection de SV est basé sur la profondeur de lectures, avec des outils tels que CNVkit s'attendant à ce que la couverture d'alignement change dans les régions avec des CNV (115). Comme ces outils ne sont pas aussi précis que GATK, l'utilisation de plusieurs méthodes en parallèle peut aider à éliminer une bonne partie des faux positifs en combinant les résultats. Des études récentes suggèrent que l'établissement du LRS sur le génome entier aidera à mieux définir la reconnaissance de SV et conduira probablement au développement d'outils de détection plus précis (82). Pour des raisons très similaires, les expansions nucléotidiques sont également difficiles à identifier. Il n'y a pas de méthode standard définie pour l'intégration aux pipelines de diagnostic, car la plupart des outils disponibles sont très récents. En effet, alors que beaucoup ont cherché à accomplir cette tâche, aucun outil ne s'est imposé comme capable de détecter de manière consistante les STR dans les données de lecture courte (116). Des chercheurs d'Illumina ont conçu ExpansionHunter (EH) pour la détection spécifique de microsatellites répétés dans 30 loci dont l'association à des maladies est bien décrite, tels que des gènes liés aux SCA, à la maladie de Huntington et aux troubles du développement neurologique (117). EH utilise un modèle de prédiction qui a été extensivement entraîné avec des données de patients où le nombre de répétitions de microsatellites est défini. Ils ont également développé une itération n'étant pas entraînée avec des STR spécifiques, mais s'efforçant plutôt d'identifier des expansions *de novo* sur l'ensemble du génome (118). Étant encore plus récent, il existe très peu de littérature concernant les performances

d'ExpansionHunter Denovo (EHdn). Les deux méthodes ont été optimisées spécifiquement pour le WGS sans PCR, mais peuvent accepter les fichiers BAM générés à partir d'autres technologies de séquençage. Similairement, STRetch est un outil développé pour des données de génome entier, mais ayant un potentiel pour l'utilisation avec le séquençage d'exome (119). Bien qu'il ne se concentre pas non plus sur des répétitions préalablement associées aux maladies, l'outil diffère d'EHdn puisqu'il effectue plutôt des tests statistiques pour chaque STR annoté sur le génome de référence. Par conséquent, il nécessite également une quantité considérable de contrôles, car il utilise une approche comparative plutôt qu'un modèle de prédiction entraîné. Si les performances sont bonnes, ces outils ont un énorme potentiel à la fois pour les tests génétiques cliniques et pour la recherche exploratoire de nouvelles STR responsables de maladies. Enfin, l'AS est un type de mutations très intéressant que l'on peut observer avec des données de séquençage transcriptomique, mais où la prédiction d'impacts fonctionnels n'est pas aussi bien définie qu'avec les SNV codants. Il existe deux approches principales pour détecter ces événements dans le RNA-seq : les méthodes basées sur les isoformes tentent de reconstruire les transcrits complets avant de les quantifier, tandis que l'approche basée sur les comptes fonctionne directement avec des lectures alignées sur les jonctions d'exons (exon-based) ou présentant un épissage (event-based), pour identifier des occurrences ou des ratios d'événements anormaux (120). Les événements d'épissage eux-mêmes sont généralement classifiés, causant soit un site donneur ou accepteur alternatif, une rétention d'intron, un saut d'exon ou une utilisation d'exons mutuellement exclusive. Ce dernier signifie qu'un seul des exons normalement co-exprimés est retenu. Plusieurs études ont démontré que l'approche basée sur les comptes surpasse son homologue (56). De plus, comme l'épissage alternatif est un processus naturel, les outils qui utilisent des statistiques comparatives ont tendance à générer des appels plus précis. Ce faisant, rMATS est l'un des outils basés sur les événements les plus populaires, puisqu'il fonctionne généralement très bien tout en étant facile à interpréter et à intégrer à un pipeline (121). DEXseq est généralement la méthode préférée pour l'analyse basée sur les exons, mais offre des performances similaires (122). Bien que l'identification des événements d'AS altérant potentiellement la fonction moléculaire soit une excellente approche pour trouver la cause d'une maladie rare, cela restreint l'identification du variant pathogène aux séquences exprimées, ce qui

n'est pas optimal pour améliorer les connaissances concernant les modèles de régulation de l'épissage alternatif. SpliceAI est un nouvel outil d'intelligence artificielle basé sur l'apprentissage profond, qui a été développé avec un réseau neural à 32 couches (66). Cela signifie que l'algorithme a été largement entraîné avec une grande quantité de variants documentés provoquant des AS afin de prédire avec précision de nouveaux événements à partir des simples changements dans la séquence nucléotidique. La principale caractéristique de cet outil est qu'il permet la détection d'un effet fonctionnel potentiel avec des données génomiques, et donc offre la possibilité d'évaluer des variants non codants éloignés des sites d'épissage (123). Bien que l'approche soit relativement nouvelle, les premières études exploitant SpliceAI suggèrent qu'il possède une précision comparable aux méthodes standard (124, 125). La précision d'appel d'épissages cryptiques pathogéniques, sites d'interaction avec le spliceosome ayant été induit par mutation, est aussi décrite (124, 126). Cependant, une limitation majeure est que la prédiction d'événement n'exploite que la séquence, et pourrait ne pas détecter un AS autrement évident. Aussi, ayant été entrainé avec des données de transcripts canoniques, SpliceAI performe moins bien pour les isoformes mineurs ou *de novo* (123). Néanmoins, l'outil pourrait avoir un impact significatif sur la génomique clinique, et est susceptible de s'améliorer en termes de performances puisque l'intelligence artificielle peut être entrainée à nouveau avec la littérature émergente.

1.6 Projet de Recherche (Hypothèse & Objectifs)

Chaque jour, les médecins diagnostiquent cliniquement de nouveaux patients atteints de maladies ayant une composante génétique, et tentent des tests géniques ciblés ou assistés par panels afin d'obtenir un diagnostic moléculaire rapide. Malheureusement, le taux de réussite de ces méthodes est généralement d'environ 20 à 50 % selon le type de pathologie (127, 128), classifiant simplement les patients restants dans la catégorie des cas complexes. Bien qu'il soit possible d'exploiter les approches NGS dans des établissements de santé, il ne s'agit pas d'une pratique courante puisqu'elles sont rarement utilisées à pleine capacité dans ce contexte en raison de contraintes de temps et de personnelles. Ces patients non diagnostiqués sont des candidats idéaux pour les laboratoires de recherche fondamentale spécialisés dans l'amélioration de l'exploration des étiologies génétiques des maladies rares, puisqu'ils sont fréquemment porteurs de variants inconnus ou VUS pouvant être ajoutés aux bases de données publiques. Le projet de recherche se concentre sur deux groupes distincts, soient une cohorte de huit patients ataxiques, et une famille avec un duo père-fille atteint par divers symptômes neurodégénératifs.

1.6.1 Cohorte du CHUM

Huit patients qui présentaient des symptômes de dégénérescence neurologique ont été référés au service de neurologie du centre hospitalier de l'Université de Montréal (CHUM : Unité des troubles du mouvement André Barbeau), où les neurologues ont effectué une évaluation clinique approfondie (Tableau III). Les phénotypes communs notables incluent l'ataxie cérébelleuse sporadique, le nystagmus ainsi que la dysarthrie, qui sont des traits typiques des ataxies épisodiques. Aucun candidat n'a été identifié dans les gènes d'ataxies lors de tests par panels, ce qui a conduit au regroupement des patients pour une étude de recherche pilote visant à identifier la cause moléculaire de leur pathologie. En effet, nous émettons l'hypothèse que la combinaison de la vue d'ensemble du génome humain offerte par le WGS, source exclusive pour les variants non codants ainsi que les SV de grandes tailles, avec les nombreuses informations fonctionnelles pouvant être extraites du RNA-seq permettra d'atteindre un rendement diagnostique supérieur aux approches individuelles.

1.6.1.1 Objectifs principaux

L'objectif principal de cette étude est d'identifier les mutations pathogéniques chez les patients, ce qui pourrait conduire à la découverte de nouvelles associations de gènes liés à l'EA, donnant possiblement une piste causative pour les formes dépourvues de caractérisation génétique. L'autre objectif majeur est de mettre en place un pipeline robuste pour l'appel ainsi que l'examen des variants nucléotidiques, comprenant l'intégration d'outils de prédiction rapides et précis pour tous les types de mutations génomiques. Par conséquent, il est important d'évaluer les performances des outils de détection STR récemment développés, ainsi que de comparer les résultats de SpliceAI avec d'autres méthodes standards d'appel d'AS. De plus, il serait intéressant d'essayer d'identifier un régulateur commun de la pathogenèse de l'EA chez plusieurs patients grâce à une méta-analyse des données transcriptomiques quantifiées. Enfin, il sera important d'évaluer l'efficacité de la démarche proposée dans cette étude, puisque la combinaison du séquençage génomique et transcriptomique entraîne un coût considérable nécessitant de bons résultats pour justifier son application future. Cela inclut l'évaluation de la pertinence des cellules mononuclées du sang périphérique (PBMC) dans l'investigation des pathologies cérébelleuse.

1.6.1.2 Impact

L'atteinte d'un diagnostic moléculaire, quoiqu'il ne se traduise pas nécessairement par l'accès à une thérapie dans le cadre des maladies rares, soulage fréquemment le patient d'une pression psychologique. En effet, non seulement cela permet au patient d'associer un nom à sa maladie, mais cela lui apporte également un sentiment d'importance aux yeux du système de santé. De plus, la confirmation d'une cause définitive améliore la capacité du clinicien à offrir des options de soins personnalisés, qu'il s'agisse de médicaments pour la gestion des symptômes traitables ou de recommandations pour une thérapie en réadaptation physique (52). L'identification de nouveaux gènes causatifs devrait permettre de mieux comprendre les événements moléculaires menant à la dégénérescence neuronale dans l'EA. Cela se traduit également par une découverte de cibles thérapeutiques potentielles pour le développement de médicaments dans un futur plus lointain. Enfin, si le taux de réussite diagnostique obtenu grâce à la combinaison de WGS et de RNA-seq est plus que respectable, l'approche aurait un potentiel bénéfique pour un éventail plus large de troubles génétiques (70).

1.6.2 Trio familial (France)

Il y a plus de dix ans, un homme a commencé à présenter des problèmes d'équilibre et de coordination, qui ont progressivement évolué vers des problèmes de mouvements oculaires, des crises pseudo-épileptiques et une myoclonie généralisée. Des analyses par résonance magnétique ont par la suite révélé que le patient présentait à la fois une atrophie cérébelleuse et une atrophie striatale bilatérale, provoquant vraisemblablement les phénotypes. Bien que plusieurs panels de gènes liés aux ataxies et aux HSP aient été effectués après des tests ciblant les gènes SCA fréquents, l'hôpital n'a pas réussi à identifier la cause génétique de la pathologie. Entre-temps, sa fille est née avec des symptômes moteurs majeurs tels qu'une incapacité à marcher, des problèmes de motricité fine, une raideur musculaire, une spasticité importante, une dysarthrie et des problèmes de mouvement des yeux. Les antécédents familiaux ont révélé que le grand-père présentait une dysphagie à apparition tardive. Cela a mené l'équipe de cliniciens à émettre l'hypothèse d'une pathologie commune causée par une expansion (31), où l'effet d'anticipation expliquerait l'aggravation des phénotypes au travers des générations. Par conséquent, la jeune fille a été testée pour les gènes d'ataxie associés aux STR qui n'avaient pas été évalués auparavant chez le père. Un séquençage d'exome a également été effectué, mais n'a donné aucun résultat et les données n'ont pas été mises à la disposition de la famille pour une investigation additionnelle. L'absence de pistes malgré la gamme de tests effectués a poussé leur établissement de santé à ne pas employer davantage de ressources pour identifier la cause de leur trouble génétique. Le raisonnement est qu'il est peu probable que cela conduise à des options de traitement étant donné les faibles bénéfices ayant résulté d'une rhizotomie dorsale chez la fille en 2015. La famille a décidé de payer pour son propre WGS de trio familial par l'entremise d'une entreprise privée avant de nous contacter pour une enquête génomique approfondie. Pour ce projet, l'hypothèse est simplement que l'utilisation d'une approche non biaisée tel que le WGS, combiné à une analyse poussée ne se concentrant pas sur un type de variant en particulier, permettra d'identifier la cause génétique de la pathologie présentée par les deux patients du trio familial.

1.6.2.1 Objectifs

Bien que les patients ne fassent pas partie du projet principal, l'objectif reste d'identifier la cause génétique de leurs troubles neurologiques. Compte tenu de l'hypothèse clinique d'une expansion STR commune, les patients devraient constituer un bon test pour les outils de prédiction, car les échantillons seront analysés via le même pipeline que les données WGS de la cohorte principale.

1.6.2.1 Impact

Encore une fois, les avantages pour les patients sont similaires en termes de gestion des soins et de soulagement psychologique. De plus, un diagnostic génétique officiel est nécessaire pour accéder à plusieurs programmes gouvernementaux de soutien financier aux personnes atteintes de maladies génétiques débilitantes, bénéficiant la jeune patiente. Finalement, il est possible que la découverte de variants chez ces patients mène à l'amélioration des connaissances moléculaires concernant le gène affecté.

2 - Méthodologie

2.1 Obtention des échantillons

2.1.1 Prélèvement sanguin

Les huit patients ont été évalués en détail par un neurologue du CHUM (Dr. Antoine Duquette) et ont subi un prélèvement sanguin dans le cadre de leur bilan clinique (Tableau III). Le comité d'éthique de la recherche du CHUM a approuvé l'étude (N°15.221). Les sujets de l'étude ont également signé un consentement éclairé autorisant l'analyse génétique dans un cadre de recherche. Le sang est prélevé dans des tubes contenant de l'acide éthylènediaminetétraacétique (EDTA) et incubé à température pièce pendant 30 minutes avant l'isolation des PBMC. Celle-ci est effectuée à par protocole standard d'isolation au Ficoll-paque (Sigma™ #17-1440-03) dans des tubes Falcon de 50mL (Sarstedt™ #62.547) pour trois cycles de centrifugation subséquents (800g, 30 minutes; 500g, 15 minutes; 200g, 10 minutes). Les cellules sont comptées suite à une dilution 1/10 sur hemacytomètre par simple microscopie en champ clair (objectif 10X). Suite à une resuspension dans un milieu de conservation (milieu de Roswell Park Memorial Institute (RPMI : Gibco™ #11875093) + 10% sérum fœtal bovin (FBS : Wisent™ #080-450) & FBS + 10% diméthyle sulfoxyde (DMSO : Sigma™ #D2650) à volume 1:1), des aliquots de 1mL contenant 20 millions de PBMC sont gelés dans des cryotubes 2mL (Sarstedt™ #72.380) et conservés dans l'azote liquide pour culture éventuelle. Des culots secs de PBMC (2.5 millions) sont également conservés à -80°C pour l'extraction des macromolécules cellulaires.

2.1.2 Prélèvement de salive

Les deux patients du trio français ont été évalués de façon exhaustive par de nombreux neurologues en France. Suite à l'épuisement des ressources hospitalières pour le diagnostic génétique, la famille s'est tournée volontairement vers le privé pour les prélèvements ainsi que le traitement d'échantillons pour séquençage. Pour les expériences de validation, des prélèvements supplémentaires ont été envoyés au CHUM. La salive était contenue dans des tubes de collection oragene (DNAgenotek™), qui ont été conservés à 4°C jusqu'à l'extraction.

Tableau III. Sommaire de la présentation clinique des patients de la cohorte.

COHORTE	MT-0007	MT-0008	MT-0009	MT-0010	MT-0011	MT-0012	MT-0013	MT-0014
Sexe	M	M	F	M	F	M	M	M
Âge	43	74	82	69	46	80	84	77
Apparition	2008	2008	2012	2007	2017	2015	2008	2004
Ethnicité	Canadien Français	Canadien Français	Italien	Français & Égyptien	Canadien Français	Tchèque	Polonais	Canadien Français
Ataxie épisodique	X	X	X	X	X	X	X	X
Progression	-[1]	X	X	X	X	X	X	X
Atrophie	Ostéophytes cervicaux	-	Cérébelleuse Temporal bila.	Cérébelleuse Cervicale	-	-	Cérébelleuse Angiome	-
Vertige	-	-	-	INI	INI	INI	-	-
Nystagmus	ACT	ACT	INI	INI	INI	INI	INI	ACT
Diplopie	INI	INI	INI	INI	-	-	INI	-
Dysarthrie	INI	INI	INI	INI	-	INI	INI	INI
Dysphagie	-	-	-	INI	-	INI	ACT	INI
Dysmétrie	-	-	INI	-	-	-	INI	ACT
Réflexes	-	Hyper (ACT)	Hyper (ACT)	Hyper (ACT)	Hyper (ACT)	Hypo (ACT)	-	Hypo (ACT)
Sensibilité altérée	Paresthésie (INI)	Vibratoire modérée (ACT)	Vibratoire légère (ACT)	Perte auditive sévère (ACT)	-	Multiple (INI)	Vibratoire sévère (ACT)	Vibratoire légère (ACT)
Tremblement	ACT	Léger (ACT)	Léger (ACT)	-	-	-	-	Léger (ACT)
Faiblesse musculaire	INI	-	INI	-	-	ACT	ACT	INI
Spasticité	-	Léger MI (ACT)	-	Modérée MI (ACT)	-	-	-	-
Autres	Céphalées Chutes	Dysfonction urinaire Péritonite	Cataractes	Aphasie Chutes Signe Babinski Vit. E élevée	Migraines Douleur oculaire	Réponse à acétalozamide Migraines Diabète	Arythmie Dyslipidémie	Chutes
Histoire familiale	Fils MT-0008	Père MT-0007 Frère similaire	Aucune	Sœur similaire	Sœur HSP	Mère possible	Père possible Mère possible	Père similaire Frère possible Fille possible

[1]Pas de progression récente; INI = Symptômes initiaux; ACT = Symptômes additionnels actuels

2.2 Extractions

2.2.1 ADN, ARN et protéines du sang

Pour chaque patient, un tube contenant 2.5 millions de PBMC a été décongelé, puis traité en suivant le protocole du kit de purification par colonne (Norgen Biotek™ #47700) pour l'extraction séquentielle de l'ADN, de l'ARN et des protéines. Une évaluation rapide par quantification spectrophotométrique a permis d'assurer la qualité de l'extraction de chacune des trois macromolécules servant aux séquençages ainsi qu'à la validation fonctionnelle subséquente.

2.2.2 ADN de la salive

L'ADN est obtenu suite à une extraction par kit (prepIT-L2P™ #PT-L2P). 0.5mL de salive est incubé à 50°C pendant 1 heure avant d'être transféré dans un tube 1.5mL pour centrifugation. Le protocole utilise le principe de précipitation de l'ADN par éthanol pure, permettant la récupération du culot par centrifugation. Après deux lavages, ce dernier est solubilisé dans 100µL de tampon Tris-EDTA (10mM Tris-Acide hydrochlorique, 1mM EDTA, pH 8.0)

2.3 Préparation des librairies de séquençage

2.3.1 Séquençage génomique

Les échantillons de la cohorte ataxie ont été dilués avec le tampon Tris-EDTA afin d'obtenir 1500ng à 2500ng dans un volume de 50µL. La plaque de séquençage Twin-Tec (Eppendorf™ #e951020401) a été envoyée au centre d'expertise et de services Génome Québec (CESGQ) pour une l'évaluation de la qualité de l'ADN, la préparation de la librairie génomique, puis le WGS. La préparation de la librairie a été effectuée par méthode « shotgun » sans PCR avec des adaptateurs « dual-index » de la compagnie Integrative DNA technologies (IDT). Le séquençage, utilisant la technologie NovaSeq 6000 S4 d'Illumina, offre une couverture de 400 millions de lectures appariées de 150 paires de bases (pb) chaque par échantillons. Le séquençage génomique effectué en France exploite également la technologie NovaSeq d'Illumina avec lectures appariées. La famille a contacté notre laboratoire afin de nous demander d'effectuer une analyse exploratoire des données, mais seuls les fichiers FASTQ ont été transmis par la compagnie privée.

2.3.2 Séquençage transcriptomique

Les échantillons de la cohorte ataxie ont été dilués avec de l'eau certifiée « RNAse-free » afin d'obtenir 600ng dans un volume de 15µL. Les tubes 1.5mL (Eppendorf™) ont été transportés sur glace sèche jusqu'au CESGQ pour une l'évaluation de la qualité de l'ARN, la préparation de la librairie génomique, puis le RNA-seq Illumina. Tous les échantillons sont évalués par bioanalyseur 2100 (Agilent™) pour s'assurer de qualité de l'ARN via une valeur d'intégrité (RIN) supérieure à 8.0. La préparation de la librairie a été effectuée par méthode de capture PolyA avec des adaptateurs « Next-dual » de la compagnie New England Biolabs (NEB). Le séquençage, utilisant la technologie NovaSeq 6000 S4 d'Illumina, offre une couverture de plus de 50 millions de lectures appariées de 100pb chaque par échantillons.

2.3.3 Séquençage Sanger

Pour chaque mutation validée par séquençage Sanger, 20µL d'amplicon (voir 2.5.1 - 2.5.3) dilué est chargé en duplicata sur une plaque Twin-Tec (Eppendorf™ #e951020401) afin de pouvoir effectuer le séquençage dans les deux directions avec les amorces correspondantes (Tableau IV), également chargé à 5uM dans une rangée distincte. Un échantillon contrôle est également toujours inclus parallèlement au patient. Une fois scellées, les plaques ont été envoyées au CESGQ pour séquençage Sanger traditionnel. Les résultats sont analysés grâce au logiciel GeneStudio.

2.4 Traitement de données

2.4.1 Alignement, contrôle qualité et annotations

Les données brutes de séquençage Illumina sont obtenues sous format FASTQ. Pour tous les échantillons génomiques, un pipeline a été mis en place afin d'utiliser l'aligneur BWA (87) sur le génome humain hg19 suite à des filtres de qualités standards par Trimmomatic (99). L'outil MultiQC permet ensuite d'obtenir un aperçu de la qualité du séquençage ainsi que de l'alignement final (97). Le pipeline d'analyse des données transcriptomiques est quasiment identique, à l'exception de l'exploitation des algorithmes STAR et HISAT2 pour l'alignement avec jonctions (91, 92). Une représentation visuelle du pipeline est disponible (Figure 3).

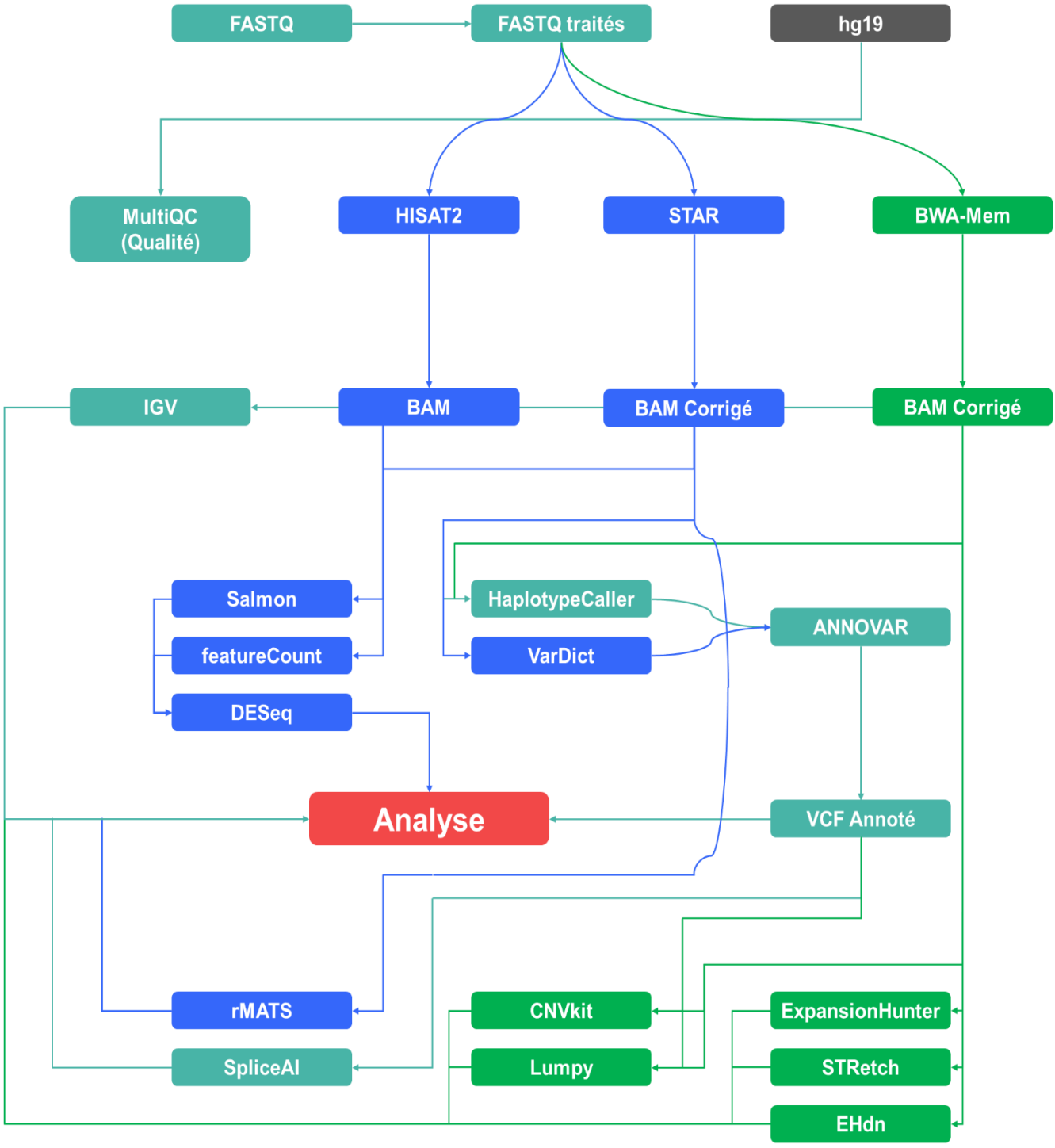

Figure 3. Schéma du pipeline bio-informatique mis en place pour le projet pilote. La couleur verte indique que le traitement est unique au WGS, la couleur bleue correspond au RNA-seq, et les processus en turquoise représentent une branche commune. Tous les outils utilisent la version hg19 du génome de référence, et certains fichiers ont été générés sans être analysés. La tête des flèches indique l'unique direction d'une étape du pipeline.

2.4.2 Outils de prédictions

Une fois l'alignement final obtenu, les fichiers BAM permettent l'appel des SNV et autres variants par GATK (v4.1.2.0), puis le fichier de sortie VCF est annoté grâce à ANNOVAR pour permettre l'analyse (101, 105). Pour les données de WGS, plusieurs outils ont été mis en place afin de prédire la présence d'AS (SpliceAI v1.2.1) (66), d'identifier les SV de petite (CNVkit v0.8.0) et grande taille (lumpy v0.3.0) (114, 115), puis de quantifier les STR déjà connues (EH v3.2.2; STRetch v0.4.0) ou prédites (EHdn v0.9.0) (117-119). Pour les données transcriptomiques, le pipeline est similaire, et l'appel de variants par Vardict (v1.6) est fait parallèlement à GATK (v4.1.2.0) pour l'évaluation des performances de ce dernier (129). Alors que les outils d'appel des SV (CNVkit et lumpy) ainsi que STRetch ne sont pas efficaces sur des données RNA-seq, l'identification d'épissage alternatif par jonctions non canoniques est possible grâce à l'outil rMATS (v4.0.2) (121). Une représentation visuelle du pipeline est disponible (Figure 3).

2.4.3 Outils de quantifications

Le pipeline de RNA-seq inclut aussi la quantification normalisée en transcrits par millions (TPM) des lectures correspondant à des gènes via l'outil Salmon pour obtenir les fichiers SF (88). Pour l'obtention de fichiers COUNT, format standard à plusieurs outils en aval de l'étape de quantification, featureCount a également été utilisé en parallèle à Salmon (130). Afin d'assurer la validité de la quantification avec HISAT2, l'alignement STAR est également quantifié par les deux outils afin de voir si les résultats étaient comparables.

2.4.4 Analyse de résultats

2.4.4.1 Évaluation complète de la littérature

L'analyse des meilleurs résultats identifier dans un VCF préfiltré pour la qualité de prédiction, les scores de pathogénicité ainsi que la fréquence allélique (AF) de population, est faites grâce à de nombreuses bases de données publiques. GeneCards regroupe l'information de plusieurs plateformes (131), et a principalement été utilisé pour déterminer la pertinence de la fonction d'un gène ainsi que de ses paralogues principaux. Les données d'expression tissulaire ont été obtenues sur ProteinAtlas (132-135), alors que l'existence d'association à des maladies connues

a pu être observée par combinaison des bases de données OMIM et Malacards (108, 131). La recherche de partenaires d'interactions possiblement reliés aux pathologies ataxiques a été faite grâce aux réseaux STRING (136). Parmi les données de cohorte utilisée pour assurer une fréquence allélique faible, puisqu'on s'intéresse particulièrement aux variants rares, gnomAD a permis de conserver uniquement les valeurs d'AF inférieures à 0.0001 (106, 137). De nombreux scores de prédictions de pathogénicité ont été attribués aux variants lorsque disponibles. Alors que les prédictions de SIFT, Polyphen2, et CADD sont considérés lors de l'évaluation de l'impact fonctionnel potentiel d'une mutation (109-111), seul ce dernier a servi à filtrer le VCF, où un seuil minimal de 15 a été mis en place pour réduire le nombre de candidats évalués en profondeur. La base de données ClinVar (43, 107), du centre national des biotechnologies d'information (NCBI), permet de savoir si une mutation a déjà été considérée dans le passé chez un patient présentant une pathologie quelconque, ou si elle n'a jamais été rapportée auparavant. Finalement, la plateforme web de gnomAD offre plusieurs informations pertinentes à l'interprétation d'un variant d'intérêt, particulièrement l'indication des valeurs d'AF de la position chromosomique pour tous les variants possible, ainsi qu'une valeur de résistance à la perte de fonction du gène (pLI) basée sur l'ensemble des données génomique du gène (106).

2.4.4.2 Visualisation intégrative

Suite à l'évaluation extensive des données filtrées, une liste ordonnée des meilleurs candidats de chaque échantillon a été obtenue. Ces derniers ont tous été validés visuellement grâce à IGV puisqu'il arrive que l'appel de variant soit parfois inexact (102), et cela pour les deux types de données. En plus de cette vérification, le RNA-seq permet d'observer la présence potentielle d'événements anormaux avoisinants le variant, tel qu'un AS via la fonction Sashimi Plot.

2.4.4.3 Compilation de données

La classification de pertinence des candidats est effectuée de façon non biaisée suite à la compilation des informations provenant de la littérature ainsi que des différents outils de prédictions bio-informatique. Ce classement définit les gènes qui seront expérimentalement évalués (Tableau IV). Dans l'optique d'une faisabilité raisonnable, la validation supplémentaire est limitée à trois gènes par patient.

2.4.4.4 Prédictions *in silico*

Certains outils de prédiction d'impact fonctionnels et structurels ont été utilisés en aval de la filtration des candidats. MutationTaster2 et I-Mutant ont servi à l'accumulation de preuves *in silico* supplémentaires pour le potentiel pathogénique des variants finaux (138, 139).

2.5 Validation de candidats

2.5.1 Conception d'amorces

Tous les candidats sont validés suite à une PCR sur le matériel nucléique disponible (ADN et/ou ARN). La conception des meilleures amorces pour l'amplification spécifique du locus d'intérêt est réalisée grâce à l'outil Primer-BLAST développé par NCBI (140). L'amplicon ciblé est toujours entre 200pb et 800pb. Toutes les séquences d'amorces utilisées sont disponibles (Tableau IV).

2.5.2 Rétro-transcription d'ARN

La transcription inverse-réaction en chaîne par polymérase (RT-PCR) est effectuée avec l'enzyme SuperScript Vilo (10X) et un mélange préconçu (5X) contenant le reste des réactifs nécessaires à la rétro-transcription (Thermo Fisher™ #11754050). Pour chaque réaction, 200ng d'ARN sont transformés en ADN complémentaire (ADNc) dans un volume 20uL. L'incubation se fait à 42°C pendant 60min, puis la réaction est arrêtée à 85°C pendant 5min et conservée à 4°C.

2.5.3 Amplification PCR

L'amplification de tous les fragments d'intérêt est faite à l'aide de 0.06U/uL d'enzyme EasyTaq, 156.3uM de dNTP, 0.15uM de chacune des amorces, ainsi que 1.15X du tampon optimisé pour la réaction (Civic Biosciences™ #AP111). Pour optimiser la réaction, 2-8% de DMSO ainsi que 0.2 à 1.7mM de sulfate de magnésium ($MgSO_4$) sont parfois ajoutés. Pour les cas complexes, 0.2U/uL de polymérase haute-fidélité Q5 avec son mélange préconçu pour les régions riches en GC sont exploités (NEB #M0491). Les conditions PCR sont standard : Dénaturation initiale (5min à 94°C); 35 cycles d'amplification (30s à 94°C, 30s à 60°C, 45s à 72°C); et élongation finale (10min à 72°C). Les fragments obtenus sont migrés sur gel d'agarose 2% (120V – 25 min) ou par système E-gel 2% (Invitrogen™ #G401002) pendant 12 minutes pour générer de plus belles photos.

Tableau IV. Variants candidats sélectionnés pour validation expérimentale

PATIENT	CANDIDAT[2]	MUTATION	AMORCES
MT-0007 & MT-0008	*ATXN7L1* (NM_138495)	Faux-sens c.2059G>A p.A687T	*ATXN7L1*_gDNA1_F: GACTCCTGTCCCCTCTCTGT *ATXN7L1*_gDNA1_R: TCGGGGAAAACGAGAACAGG *ATXN7L1*_cDNA1_F: AATGCTGTGTCTTCTCTGCCC *ATXN7L1*_cDNA1_R: TTATCCTGCCCGTTCTGTTTG
MT-0007 & MT-0008	*SEC14L6* (NM_001193336)	Faux-sens c.1021C>A p.R341S	*SEC14L6*_gD1_F: CGCCAATGGGGCTCTATCCA *SEC14L6*_gD1_R: GTAGGACCTGGATGCACATTGTA *SEC14L6*_cD1_F: TCAGGTGGCAGTTTGCTTCA *SEC14L6*_cD1_R: CACGGTGTAGCTGATGCGTT
MT-0008	*KCNA4* (NM_002233)	Épissage c.-783+5G>T	*KCNA4*_gD1_F: GCCAAACCCGAGTGATTCCT *KCNA4*_gD1_R: CCCTTCTGTCACATTCCCGA *KCNA4*_cD1_F: TTTTCGGGGGAACCTTGACT *KCNA4*_cD1_R: ATGAGGTGTCAGCAGAAGCAA
MT-0008	*VPS13C* (NM_017684)	Faux-sens c.11014G>T p.D3672Y Épissage c.4165+1192A>C	*VPS13C*_gD1_F: ACCGTCCTTGTTCCAAGATGAT *VPS13C*_gD1_R: TAAGACCCTTGCACCCGACT *VPS13C*_cD1_F: AGTTGGAAGGAGAGACTTACCG *VPS13C*_cD1_R: TCTCGTTGACTGTGCATCCTC *VPS13C*_gD2_F: TTGACAGATCTTGCCGATTG *VPS13C*_gD2_R: CGAATGAATCTCGCAAACAA *VPS13C*_cD2_F: GCAGTCACTTTTTGCCCGAC *VPS13C*_cD2_R: ACACGAAAACTCCTCTGGGC
MT-0009	*ELOVL4* (NM_022726)	Épissage c.541+5G>A	*ELOVL4*_gDNA1_F: AAGGAGTTGAGTATTTGGACACA *ELOVL4*_gDNA1_R: CAAAGTCCTAGGTTCTCATTGCT *ELOVL4*_cDNA1_F: TCTGATGCAGTCTCCTTGGC *ELOVL4*_cDNA1_R: GGAAGGGGCAGTCAGTGTAA
MT-0009	*KCNAB3* (NM_004732)	Faux-sens c.1009G>A p.A337T	*KCNAB3*_gD1_F: GGCGGGGTCTTCTGTTGATTC *KCNAB3*_gD1_R: GTCTCTGAAGACAAGCGTCC *KCNAB3*_cD1_F: AGCAAGTATGATGGGCGAGTC *KCNAB3*_cD1_R: CGGAGACACCACGCAATAG
MT-0010	*PMPCB* (NM_004279)	Épissage c.1154+5G>C	*PMPCB*_gDNA1_F: AAGCGTATGTAGCCAAGAGTCC *PMPCB*_gDNA1_R: GGAAAAACCAACTGCAACCTTTG *PMPCB*_cDNA1_F: CTCTGCCTCCCTGCAAATTC *PMPCB*_cDNA1_R: GGGACCAACAGCAGCAATAG
MT-0010	*MARS* (NM_004990)	Faux-sens c.2210G>A p.R737Q	*MARS*_gD1_F: CCTCTGTCTTTTCTGCTGGC *MARS*_gD1_R: GGTCTTCTCTCACCCACTCC *MARS*_cD1_F: TATGTGCCTGAGATGGTGCT *MARS*_cD1_R: TCTGGTGTCCTGCTGGTAAG
MT-0010	*STAC2* (NM_198993)	Épissage c.496-3T>G	*STAC2*_gD1_F: GGGAAGGACCAAGTTCTCCA *STAC2*_gD1_R: ACCCCAAATTCAGGACAGGA *STAC2*_cD1_F: TGCCACCAGCTCATCGTAG *STAC2*_cD1_R: GGGGCTGTGAATACTGGAGA

MT-0011	*GABRP* (NM_001291985)	Faux-sens c.445C>A p.L149M	*GABRP*_gDNA1_F: TTTCTCACCCACTACCCCCA *GABRP*_gDNA1_R: CCCCCAAACCACTCACCTAC *GABRP*_cDNA1_F: TTTGGTGGAGAACCCGTACA *GABRP*_cDNA1_R: CACAGAGTCGTTCCCTCTCA
MT-0012	*SPG7* (NM_003119)	Non-sens c.1861C>T p.Q621X Faux-sens c.2228T>C p.I743T	*SPG7*_gDNA1_F: TGCCTTCCTGCTTTGAGACG *SPG7*_gDNA1_R: TGCACTGGAACAGAAGGAGTC *SPG7*_gDNA2_F: TGAGGTTGAGATGGGGGTGA *SPG7*_gDNA2_R: TCGCCCAAGTCCTGTTTCTC *SPG7*_cDNA1_F: AGAACAGAAAGTGGTTGCGT *SPG7*_cDNA1_R: CCCAAGTCCTGTTTCTCCCT
MT-0012	*ATXN7L1* (NM_138495)	Faux-sens c.40A>T p.N14Y	*ATXN7L1*_gD2_F: CCGCCTGTTGTGTGTTTGAG *ATXN7L1*_gD2_R: ATGGTGACAGGAACAGCAGG *ATXN7L1*_cD2_F: CGTTCTCGAATCCCGTGTCT *ATXN7L1*_cD2_R: ATTGGCACCATCTGCCTTCA
MT-0012	*ARHGAP4* (NM_001666)	Faux-sens c.2294T>C p.L765P	*ARHGAP4*_gD1_F: ACAGCTTGGACCCCCTCT *ARHGAP4*_gD1_R: GGTGGATGAAGGGGAGACT *ARHGAP4*_cD1_F: CGTCTACGAGAAGTGCATGG *ARHGAP4*_cD1_R: AGGTGAGGTGCATGGCTCT
MT-0013	*ZFYVE26* (NM_015346)	Non-sens c.3022G>A p.R1008X	*ZFYVE26*_gDNA1_F: TGAAGGCAGTACCAAGGCAA *ZFYVE26*_gDNA1_R: GCTGACCTAATGTTCCAAGTCC *ZFYVE26*_cDNA1_F: GAACTCAGATGCGGGTAGCA *ZFYVE26*_cDNA1_R: GGCAACACAGTCCTCGCTTA *ZFYVE26*_cD2_F[1]: AGTTGCCTGCTTGTACTCCT *ZFYVE26*_cD2_R[1]: AGTCATTCGCTGGCTCTGTA
MT-0013	*ATXN2* (NM_002973)	Expansion (CAG_{32})	*ATXN2*_gDNA1_F: CTTCGTCGTCCTCCTTCTCC *ATXN2*_gDNA1_R: TCCCTCCATCTTGACCGC
MT-0014	*CACNA1H* (NM_001005407)	Faux-sens c.4772G>A p.R1591Q Faux-sens c.2354A>T p.K785M	*CACNA1H*_gD1_F: ACCCATGACACCTGCAAAGAT *CACNA1H*_gD1_R: CCTGGTCTCCTATCCTCACCA *CACNA1H*_gD2_F: AGAGTCCCCCTTCTCCAGTC *CACNA1H*_gD2_R: AAAAGGCTGGGAGGACGAAG *CACNA1H*_cD1_F: CTACGAGAAGATCCCGCATGT *CACNA1H*_cD1_R: AGCATTAGTCAGCTCCTCGG *CACNA1H*_cD2_F: ACATCTCCACCAAGGCACAG *CACNA1H*_cD2_R: CGGCGCTCATCTCTATCTCC
Fille	*SPAST* (NM_014946)	Faux-sens c.1400G>A p.R499H	SPG4_gD1_F: AGCTTTTCCTGTCATTTGCTG SPG4_gD1_R: GCTGTACCATGGATTGGAAGA
Père	*RFC1* (NM_002913)	Expansion ($AAAAG_n$)	*RFC1*_gD1_F: TCACGCCTGTAATCCCAGCATTG *RFC1*_gD1_R: TCTTGAAGAATAGCTGTGTTGTCCTGTCAC

[1]Recherche d'événement d'épissage; [2] Seul l'isoforme de référence pour mutation est indiqué.

2.5.4 PCR quantitative (qPCR)

10ng/uL d'ADNc sont mélangés avec une amorce Taqman (20X) spécifique au gène d'intérêt et un mélange Taqman optimisé (2X) pour l'obtention de signaux FAM-MGB dans une plaque 384 puits (Thermo Fisher™ #AB1384) avec un volume final de 10uL par puits. L'information sur les amorces utilisées est disponible (Tableau V). La β-Actine est utilisée comme gène de référence pour normaliser l'expression dans les cellules PBMC, alors que GAPDH est une référence plus standard pour la salive (141, 142). La plaque est centrifugée rapidement puis l'amplification par PCR quantitative ainsi que la comparaison des signaux fluorescents en mode relatif (ΔΔCT) est réalisée par le système QuantStudio 6/7 (Applied Biosystems™). Les quantifications relatives sont normalisées en fonction du groupe de contrôles sains, et une ANOVA non-paramétrique avec correction post-hoc de type Tukey HSD est utilisée comme test statistique.

Tableau V. Amorces pour qPCR des gènes candidats

PATIENT	CANDIDAT	LOCALISATION	AMORCES
MT-0007	*ATXN7L1*	Jonction exonique 7-8	Hs00393420_m1
MT-0008	*SEC14L6*	Jonction exonique 8-9	Hs01596536_m1
MT-0009	*ELOVL4*	Jonction exonique 1-2	Hs00224122_m1
MT-0010	*PMPCB*	Jonction exonique 4-5	Hs00188704_m1
MT-0013	*ZFYVE26*	Jonction exonique 16-17	Hs01012489_m1
France-MT-F	*RFC1*	Jonction exonique 15-16	Hs01099126_m1
Contrôle PBMC	B-Actine	Jonction exonique 2-3	Hs01060665_g1
Contrôle salive	GAPDH	Exon 7	Hs002786624_g1

3 - Résultats

3.1 Validation de la pertinence du sang pour les projets d'ataxie

À la suite d'un recensement des gènes associés aux diverses formes de maladies impliquant des phénotypes ataxiques primaires ou secondaires (Tableau II), l'utilisation de multiples bases de données a permis de prédire le potentiel de détection de mutations pertinentes dans le RNA-seq. L'Atlas de protéine humaine (HPA), source principale de l'information fonctionnelle recueillie, est une plateforme regroupant plusieurs jeux de données transcriptomiques et protéiques tels que GTEx ou FANTOM5 (134, 135). Les bases de données GeneCards, BioGPS, ainsi que la plateforme de l'institut européen de bio-informatique (EBI) ont également contribué à l'évaluation de l'expression génique des cellules du sang (Figure 4), mais toutes les données sont combinées afin d'illustrer le consensus pour chaque gène (131, 133, 143). Au niveau de l'ARNm, c'est 157 des 169 gènes (92.8%) qui sont supposément exprimés dans les cellules sanguines (Figure 4A), avec une difficulté un peu plus grande de détection des gènes causatifs d'ataxies primaires AD où un peu plus de 14% ne sont habituellement pas détectés. D'un point de vue plus nuancé, près de 80% des causes connues d'ataxies devraient présenter des niveaux d'expression assez importants pour détecter la présence d'altération fonctionnelle (135/169), alors que l'analyse devrait se limiter à la découverte de variants dans près de 15% des gènes puisque les niveaux attendus sont trop faibles pour une comparaison statistique (Figure 4B). Suite à la réception des données NovaSeq, la quantification a permis de confirmer que c'est en fait 84% des gènes (142/169) qui offrent la possibilité d'être évalué à un niveau statistiquement significatif (Figure 4E-F). Cependant, plus de 11% ne sont pas détectés du tout (20/169), avec la même tendance où les causes d'ataxies AD sont plus difficiles à observer (25/33) que les formes AR (62/70) ainsi que secondaires (62/66). L'évaluation des niveaux protéiques théoriques a également été faite (Figure 4C-D), suggérant un potentiel pour la validation fonctionnelle de futurs candidats chez 62% de gènes (105/169). À noter, aucune donnée protéique n'est disponible pour un nombre considérable de causes d'ataxies (26/169), les quantifications se limitent parfois à un type de lymphocyte spécifique, et un niveau cérébelleux nul peut mener à une classification « similaire » sans expression sanguine.

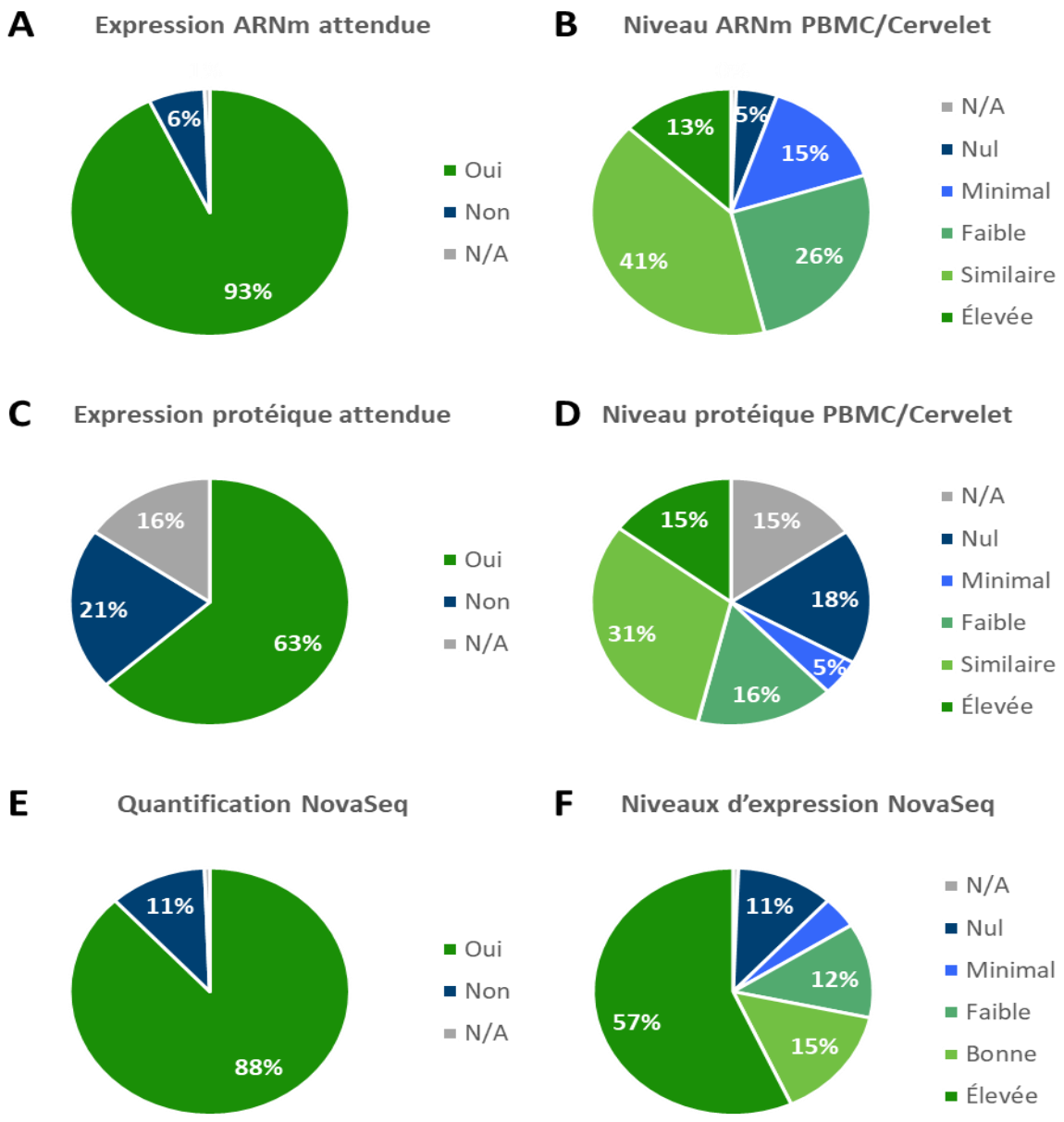

Figure 4. Niveaux d'expression des gènes d'ataxies dans les PBMC sanguins. Présence consensus d'ARNm dans les cellules sanguines (**A**) selon les bases de données et niveaux observables en comparaison avec le cervelet (**B**). Présence consensus de protéines dans les cellules sanguines (**C**) selon les bases de données et niveaux observables en comparaison avec le cervelet (**D**). Détection (**E**) et interprétation qualitative des niveaux d'ARNm (**F**) présents dans les données NovaSeq. Les catégories sont définies en fonctions du nombre de lectures, comme suit : Nul = 0-2; Minimal = 3-20; Faible = 21-200; Bonne = 201-800; Élevée = 800+; N/A = Non-applicable (pour locus chromosomique non génique ou absence de données).

3.2 Évaluation de performances d'outils bio-informatiques

Étant donné la mise en place du pipeline bio-informatique pour le projet pilote, la majorité des outils ont été installés et utilisés pour la première fois avec les données de la cohorte EA. Cependant, seules les performances des algorithmes de prédictions STR ainsi que de SpliceAI étaient mises en question due à leur nouveauté dans la littérature (66, 117, 119). Il est cependant à noter que l'appel de variant effectué par HaplotypeCaller sur les données RNA-seq a mené à un nombre anormalement élevé de faux positifs (non présenté). Le reste des pipelines d'analyse suggèrent des performances correspondant à ce qui est prévu dans la littérature.

3.2.1 EH & STRetch

Afin d'évaluer la précision des outils d'appel de STR, ExpansionHunter et STRetch, les données WGS de 54 contrôles négatifs ainsi que 34 contrôles positifs ont été téléchargées sur des bases de données publiques. Parmi ces derniers se trouvent des porteurs d'expansions *FXN* (10), *FMR1* (5), *DMPK* (4), *ATXN1* (4), *C9ORF72* (4), *HTT* (4), *AR* (1), *ATN1* (1), ainsi qu'*ATXN3* (1) provenant de deux projets différents. Les données négatives quant à elle proviennent d'une source distincte, mais il est important de noter que contrairement à EH, STRetch nécessite ces contrôles lors de son entraînement, et donc ceux-ci ne sont pas inclus dans les statistiques de spécificité. Un appel positif pour une expansion de STR est défini comme une prédiction du nombre de répétitions nucléotidiques supérieur au seuil pathogénique (Figure 5A) défini dans la littérature (32, 144). À l'opposé, un faux positif est défini par l'appel d'une expansion pathogénique en fonction de ce même seuil pour un contrôle sain. Les performances d'EH (Figure 5B) démontrent une sensibilité de détection de 94.1% (32/34) comparativement à seulement 64.7% pour STRetch (22/34). Quoique les deux outils tendent à sous-estimer le nombre de STR, la précision d'EH est supérieur, se rapprochant généralement davantage du nombre réel. Au niveau de la sensibilité (Figure 5B), EH induit fréquemment des faux positifs pour les expansions dans *NIPA1*, *PHOX2B*, ainsi que *TCF4*. En ignorant seulement ces prédictions, l'outil offre une spécificité ajustée de 96.3% (52/54) comparativement à 75.0% (3/4) pour STRetch. Les performances d'EHdn quant à elle étaient très faibles pour les expansions connues, suggérant que la méthode d'évaluation n'est pas appropriée à l'algorithme, et expliquant son absence dans la comparaison.

A

Gènes évalués	Microsatellite	Seuil pathogénique
FXN	GAA	> 34-65
FMR1	CGG	> 55-199
DMPK	CAG	> 35-49
ATXN1	TGC	> 36-38
C9ORF72	GGCCCC	> 26-59
HTT	CAG	> 27-39
AR	GCA	> 35-38
ATN1	CAG	> 36-47
ATXN3	GCT	> 45-59

B

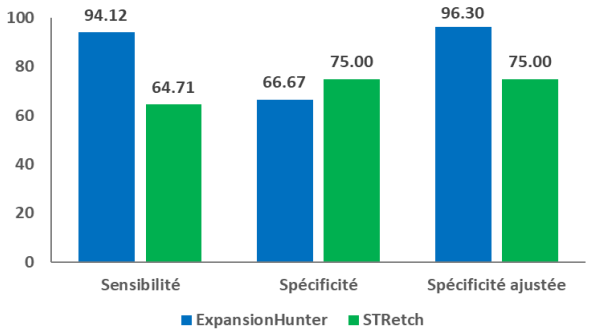

Figure 5. Performances des outils de prédictions STR. Définition des gènes évalués ainsi que du seuil de pathogénicité utilisé pour définir la précision des appels (**A**). Spécificité et sensibilité des outils EH ainsi que STRetch, défini à partir de la précision d'appels sur contrôles positifs (n = 34) et négatifs (n = 54 pour EH; n = 4 pour STRetch) obtenues de base de données publiques (**B**).

3.2.2 SpliceAI

Une analyse qualitative des résultats obtenus permet de déterminer la pertinence de l'intégration de l'outil dans un pipeline clinique (Tableau VI). Les données générées contiennent, pour chaque variant nucléotidique, au moins quatre scores d'impact sur l'épissage. Ceux-ci sont définis comme la perte (L) ou le gain (G) d'un site donneur (DG; DL) ou accepteur (AG; AL) (66), et cela pour différents types de mutations. Dans l'analyse (Tableau VI), tous les variants se trouvant sur une séquence traduite sont dits codants, alors que les variants non codants sont séparés en fonction de leur positionnement comparativement au gène le plus près. Tel qu'indiqué, certains se situent dans les régions introniques et dans les séquences non traduites (UTR) de l'ARNm, alors que « autres » regroupe des variants en amont et en aval de gènes. Seules les mutations ayant au moins 1/4 score prédit au-deçà de 0.20 sont conservés et quantifiés suite à l'appel, tandis que les scores supérieurs à 0.85, considérés comme ayant un potentiel clinique (124), ont également été quantifiés séparément. Alors que les données de WGS contiennent généralement près de 400 événements prédits, dont une vingtaine offrent un score cliniquement pertinent, 2.1% à 7.6% de ce nombre d'AS sont présent dans le RNA-seq. Une bonne partie de ces résultats concordent, mais un nombre proportionnellement faible de résultats transcriptomiques ont un score supérieur à 0.85. En moyenne, un peu moins de 10% des événements sont près des jonctions d'exons, région où la conservation de séquence est généralement considérée comme essentielle à l'épissage canonique (76, 77).

Tableau VI. Évaluation des performances de SpliceAI

Analyse des données génomiques							
Échantillons	Codant	Intronique	Autres non-codant	Région non-traduite	Total	Jonction d'épissage	Score > 0.85
MT-0007	46	185	125	21	377	27	25
MT-0008	39	184	136	20	379	29	21
MT-0009	38	196	147	30	411	47	23
MT-0010	31	159	131	28	349	31	14
MT-0011	48	179	133	23	383	38	19
MT-0012	38	204	149	25	416	34	17
MT-0013	38	181	136	20	375	34	22
MT-0014	36	175	137	29	377	38	18
MT-0015	55	181	141	28	405	21	19
Analyse des données transcriptomiques							
Échantillons	Codant	Intronique	Autres non-codant	Région non-traduite	Total	Jonction d'épissage	Score > 0.85
MT-0007	4	1	3	0	8	0	0
MT-0008	14	3	5	1	23	0	2
MT-0009	8	3	9	1	21	0	1
MT-0010	15	3	5	3	26	2	0
MT-0011	12	2	11	1	26	0	0
MT-0012	14	3	8	1	26	1	0
MT-0013	15	2	7	3	27	0	0
MT-0014	16	5	5	0	26	1	2
MT-0015	16	2	10	3	31	1	0

Chaque valeur correspond au nombre de variants avec cette caractéristique.

3.3 Identification de variants candidats

L'appel de variant par GATK est central aux pipelines d'analyse génomique et transcriptomique. La majorité (> 99.92%) des variants détectés dans le VCF initial (Tableau VII) sont éliminés par les filtres standards aux meilleures pratiques d'analyse de données WGS (83, 84, 100). Un filtrage additionnel en fonction des fréquences alléliques (gnomAD ≤ 0.0001) ainsi que des conséquences fonctionnelles prédites par CADD (≥ 15) a réduit le nombre de variants évalué manuellement à plus ou moins de 3% de la quantité du rapport standard. En excluant des filtres manuels les variants se trouvant dans des gènes préalablement associés aux ataxies, il est possible d'obtenir un peu plus de 6% des résultats correspondant à cette catégorie en moyenne. De ceux-ci, le nombre de variants étant considéré comme candidat potentiel est également dénoté, avec seulement un patient démontrant de bons candidats qui auraient dû être identifiés via les tests traditionnels par panel ou séquençage ciblé. L'appel par HaplotypeCaller sur le RNA-seq contenait plus de 70% de faux positifs dû à une erreur non résolue, mais rapportée dans la communauté pour cet outil. La mise en place subséquente de l'outil Vardict avec une suite de pipeline identique (Figure 3) a donc permis d'obtenir les résultats présentés. Outre le fait que les filtres standard ne semblent éliminer que ≈97% des appels ici, les statistiques sont similaires aux données WGS.

Tableau VII. Statistiques concernant le traitement des données d'appel de variants

Analyse des données génomiques (HaplotypeCaller)						
Échantillons	**Total**	**Filtres Standards**	**Filtres Manuels**	**Gène d'ataxie**	**Candidat ataxie**	**NGS essentiel**
MT-0007	4707739	3366	108	9	Non	Oui
MT-0008	4688777	3336	120	4	Non	Oui
MT-0009	4750038	3619	136	6	Oui (1)	Oui
MT-0010	4708157	3339	98	9	Oui (1)	Oui
MT-0011	4735119	3391	106	6	Oui (1)	Oui
MT-0012	4732700	3615	146	12	Oui (2)	Non
MT-0013	4706386	3302	95	9	Non	Oui
MT-0014	4826787	3507	86	7	Non	Oui
MT-0015	4830684	3677	111	5	N/A	N/A
Analyse des données transcriptomiques (Vardict)						
Échantillons	**Total**	**Filtres Standards**	**Filtres Manuels**	**Gène d'ataxie**	**Candidat ataxie**	**NGS essentiel**
MT-0007	165394	7912	247	17	Non	Oui
MT-0008	155757	7227	226	13	Non	Oui
MT-0009	184816	7753	221	11	Non	Oui
MT-0010	313272	9236	342	16	Non	Oui
MT-0011	247523	8547	282	24	Non	Oui
MT-0012	305334	9225	342	19	Oui (2)	Non
MT-0013	244167	9148	279	16	Non	Oui
MT-0014	257164	8990	317	14	Non	Oui
MT-0015	289307	8867	307	21	N/A	N/A

Chaque valeur correspond au nombre de variants restant après chaque filtre.

N/A = Non-applicable (Contrôle en bonne santé)

3.3.1 Cohorte EA

Pour chaque patient de la cohorte, les données WGS ont permis d'obtenir des prédictions de variants pathogéniques à partir des outils GATK, SpliceAI, EH, EHdn, STRetch, CNVkit ainsi que Lumpy (Figure 3). À moins d'une identification de cause pathogénique évidente, la totalité des résultats génomiques a été analysée (Section 2.4), et les variants ayant une implication potentielle dans les phénotypes ont été classifiés afin d'obtenir les meilleurs candidats pour chaque patient. Ceux-ci sont présentés (Tableau VIII) conjointement à la pathologie qui leur est associée, le type de variant concerné, ainsi que l'outil ayant permis la détection. Pour les données RNA-seq, des appels ont été générés avec les outils GATK, Vardict, rMATS en plus des quantifications pour expression différentielle par Salmon et featureCounts. Cependant, à la suite de la détection d'un nombre anormalement élevé de faux positifs pour GATK, les résultats n'ont été utilisés que pour valider les candidats du WGS ou bien pour identifier des conséquences fonctionnelles reliées aux mutations d'intérêts. Cela signifie que les variants uniques aux données transcriptomiques n'ont pas encore été analysés et incluent dans la recherche de candidats présentés (Tableau VIII).

Tableau VIII. Meilleurs gènes candidats identifiés pour les patients de la cohorte EA.

	Candidat	Maladie associée	Type variant	AF gnomAD	Score CADD	Tolérance pLI	Classification	Outil	Expression sang
MT-0007 & 08	ATXN7L1*	SCA7[1] (AD)	SNV faux-sens	$6.56x10^{-5}$	23.8	1	VUS	GATK	Oui
MT-0007 & 08	SEC14L6*	AVED[1] (AR)	SNV faux-sens	$1.19x10^{-5}$	24	0	VUS	GATK	Non
MT-0008	KCNA4	EA1[1] (AD)	Épissage alternatif	0	0	0.8	VUS	SpliceAI	Non
MT-0008	VPS13C	Parkinson 23 (AR)	2 x SNV faux-sens	$9.51x10^{-4}$ / $2.74x10^{-3}$	26.6 / 0	0	2 x VUS	GATK; SpliceAI	Oui
MT-0009	ELOVL4*	SCA34 (AD)	Épissage alternatif	0	19.41	0.83	VUS	GATK; SpliceAI	Oui
MT-0009	KCNAB3	-	SNV faux-sens	$1.11x10^{-5}$	24.5	0	VUS	GATK	Oui
MT-0010	PMPCB*	MMDS6[2] (AR)	Épissage alternatif	$4.91x10^{-5}$	16.7	0	VUS	GATK; SpliceAI	Oui
MT-0010	MARS	CMTD-2U[3] (AD)	SNV faux-sens	0	34	0	VUS	GATK	Oui
MT-0010	STAC2	Myopathie (AR)	Épissage alternatif	$4.55x10^{-6}$	9.51	0	VUS	SpliceAI	Non
MT-0011	GABRP*	EA3[1]-EA4[1] (AD)	SNV faux-sens	$1.19x10^{-5}$	25.7	0	VUS	GATK	Non
MT-0012	SPG7*	HSP7 (AD/AR)	Non-sens + faux-sens	0 / $4.95x10^{-5}$	39 / 25.4	0	Path. / Prob. Path.	GATK	Oui
MT-0012	ARHGAP4	Diabète (AD)	SNV faux-sens	0	24.5	0.98	VUS	GATK	Oui
MT-0012	ATXN7L1	SCA7[1] (AD)	SNV faux-sens	0	17.4	1	VUS	GATK	Oui
MT-0013	ATXN2*	SCA2 (AD)	Expansion	-	0	0.85	Pathogénique	EH	Oui
MT-0013	ZFYVE26*	HSP15 (AD/AR)	Non-sens	0	42	0	VUS	GATK	Oui
MT-0014	CACNA1H*	Épilepsie (AR)	2 x SNV faux-sens	$3.53x10^{-3}$ / $3.37x10^{-4}$	24.2 / 27.5	0	Bénin / VUS	GATK	Oui

[1]Association par inférence automatique de la littérature; [2]MMDS = Syndrome de dysfonctions mitochondriales multiples; [3]Maladie de Charcot Marie-Tooth Type 2U; *Candidat principal du patient

3.3.2 Trio familial

Puisque les données de séquençage ont été générées avec une technologie similaire au WGS de la cohorte ataxie, les échantillons du trio français ont été soumis au même pipeline génomique non biaisé contenant les outils GATK, SpliceAI, EH, EHdn, STRetch, CNVkit ainsi que Lumpy. L'appel de variant effectué avec HaplotypeCaller de GATK a permis de rapidement identifier un SNV faux-sens dans le gène *SPAST*, aussi connu sous le nom de *SPG4* (c.1496G>A; p.R499H), chez la jeune fille. Ce variant a précédemment été rapporté comme « probablement pathogénique » dans la base de données ClinVar (43). Malgré le fait qu'il s'agisse d'une mutation *de novo* n'étant pas partagé avec son père, le variant est immédiatement devenu l'unique candidat sélectionné pour expliquer le désordre ataxique et paraplégique à présentation précoce. Pour le père, c'est plutôt l'outil EH qui a permis d'identifier le meilleur candidat pour expliquer la pathologie cérébelleuse à présentation tardive. L'outil prédit une expansion du STR-AAAAG contenu dans l'intron 2 de la sous-unité 1 du facteur de réplication C (*RFC1*), répétition associée au syndrome d'ataxie cérébelleuse, neuropathie, et aréflexie vestibulaire (CANVAS) (145). Cependant, cette expansion n'est pas détectée par STRetch ou EHdn, puis le nombre de répétitions prédit par EH n'est que 37 et 43 pour les deux allèles respectifs, ce qui n'est pas suffisant pour établir un diagnostic.

3.4 Validation expérimentale des candidats

3.4.1 Validation de la cohorte EA

Au total, 20 variants ont été sélectionnés suite à l'analyse des résultats du pipeline WGS pour une validation expérimentale, cette dernière ayant permis de définir les 11 meilleurs candidats causatifs pour les huit patients du projet (Tableau VIII). L'observation des données RNA-seq a permis de confirmer la présence des variants dans tous les gènes exprimés dans le sang (16/20), préalable à la vérification des SNV par séquençage Sanger. Pour les meilleurs candidats ayant un niveau d'expression acceptable dans les PBMC, une qPCR a subséquemment été effectuée si une validation supplémentaire s'avérait nécessaire à la corrélation du génotype au phénotype. Les échantillons sains MT-0001 à MT-0003 ainsi que MT-0015 sont utilisés de façon consistante pour contrôler les résultats de validation de la cohorte EA. Alors que ce dernier a été traité exactement de la même façon que MT-0007 à MT-0014 durant le projet, MT-0001 à MT-0003 ont été prélevés et extraits en utilisant la même méthodologie, mais quelques mois avant le début du projet.

3.4.1.1 Variants partagés dans le duo MT-0007 & MT-0008

Les deux meilleurs candidats du duo père-fils, soient *ATXN7L1* (c.2059G>A;p.A687T) et *SEC14L6* (c.1021C>A;p.R341S), ont subi une validation longue (Figure 6) puisqu'ils ne sont associés à des pathologies ataxiques que par inférence. Suite à une PCR amplifiant la région d'intérêt de chacun des gènes (Figure 6A; 6D), le séquençage Sanger des fragments d'ADNg et d'ADNc dans les deux directions a été effectué; puisque les résultats corrèlent parfaitement, un seul des quatre séquençages est présenté par patient dans le but d'être concis (Figure 6B; 6E). Les deux variant sont confirmés comme étant hétérozygotes dans les patients. La quantification par qPCR a été réalisée en triplicata avec les trois contrôles extra-cohorte (MT-0001 - MT-0003) ainsi qu'avec le contrôle intra-cohorte (MT-0015). Les signaux Taqman sont analysés en mode relatif delta delta Ct (cycle dépassant le seuil de fluorescence). Le patient MT-0012 est inclus dans la quantification de l'ARNm d'*ATXN7L1* puisqu'il possède un VUS distinct sur le gène (c.40A>T:p.N14Y). Le duo présente une diminution négligeable d'*ATXN7L1*, étant plutôt absente pour MT-0012 (Figure 6B). Pour *SEC14L6*, les niveaux d'ARNm observés sont plus similaires entre les deux patients, mais la diminution demeure non significative selon l'ANOVA non-paramétrique (Figure 6E).

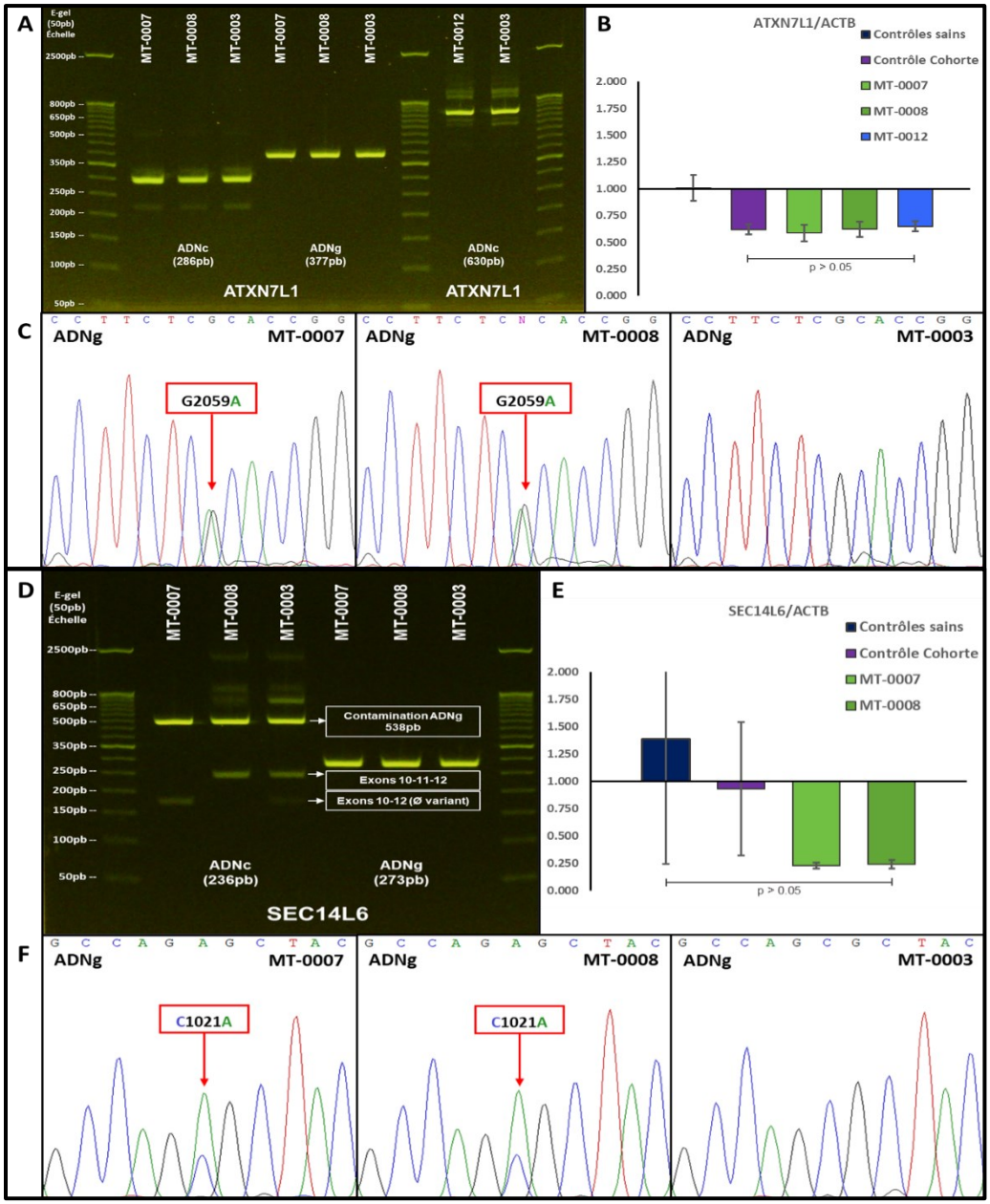

Figure 6. Validation expérimentale des variants *ATXN7L1* et *SEC14L6* chez le duo père-fils. Les régions flanquant les variants dans l'ADN ainsi que dans l'ARNm sont amplifiées par PCR et les fragments résultants sont migrés sur E-gel 2% pour assurer la qualité des échantillons (**A**, **D**). Ceux-ci sont séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à un contrôle (**C**, **F**). Dans une plaque 384 puits, un triplicata contenant 15ng d'ADNc est incubé avec une sonde TaqMan pour chaque gène d'intérêt. La quantification du signal fluorescent pas le système QuantStudio 6/7 est analysé en mode ΔΔCt relativement à l'expression de l'actine-β (**B**, **E**). Les contrôles sont groupés selon la date d'extraction de l'ARNm.

3.4.1.2 MT-0008

Un variant UTR causant potentiellement un épissage alternatif du gène *KCNA4* (c.-783+5G>T) ainsi que duo de mutations SNV faux-sens (c.11014G>T;p.D3672Y) et d'épissage (c.4165+1192A>C) dans le gène *VPS13C* avaient également été identifiés comme candidats étant unique à MT-0008. Alors qu'il était possible de voir les trois variants dans les données de WGS et RNA-seq, l'amplification par PCR de l'ADNc de *KCNA4* dans le but d'évaluer la présence d'un événement d'épissage n'a pas été possible (non présenté), disqualifiant le candidat d'une validation plus approfondie. Similairement, l'amplification de l'ADNc de *VPS13C* a permis d'observer un épissage en apparence normal (non présenté), mettant fin au processus d'évaluation de ce gène.

3.4.1.3 MT-0009

Nonobstant le potentiel majeur du variant intronique identifié dans *ELOVL4* (c.541+5G>A), un SNV faux-sens observé dans le gène *KCNAB3* (c.1009G>A;p.A337T) a également été sélectionné pour une possible validation expérimentale. Malgré la confirmation du variant par séquençage Sanger de l'ADNg et ADNc amplifiés (non présenté), l'évaluation de ce candidat secondaire a été mise de côté suite aux résultats obtenus pour *ELOVL4*. Dès la vérification de l'amplification sur gel d'agarose (Figure 7A), il est possible de voir un profil d'isoformes divergent de celui du contrôle dans l'ADNc. En effet, deux bandes supplémentaires sont clairement visibles, soit des fragments 436pb et 308pb spécifiquement. Suite au séquençage Sanger (Figure 7B), il est possible d'interpréter ces bandes comme des nouveaux transcrits démontrant un AS qui engendre le saut de l'exon 4, puis de l'exon 4-5 respectivement. En effet, les signaux obtenus démontrent la présence des jonctions non canoniques des exons 3-5, ainsi que 3-6, qui n'apparaissent pas du tout dans le contrôle. Ce dernier confirme également la présence négligeable d'artéfacts de séquençage présent indépendamment du variant d'intérêt. La vérification de la présence du variant intronique, chez MT-0009, mais pas MT-0001, a bien évidemment été réalisée avec l'ADNg (Figure 7B). La quantification de l'ARNm *ELOVL4* dans les PBMC (Figure 7C) démontre une diminution importante de l'expression du gène dans MT-0009 comparativement aux contrôles intra et extra-cohorte. La quantification des résultats préliminaires suggère une diminution des niveaux d'*ELOVL4* de près de 40% et 60% respectivement. Cette différence est non significative selon l'ANOVA non-paramétrique avec correction Tukey ($p > 0.05$). L'expérience sera répliquée.

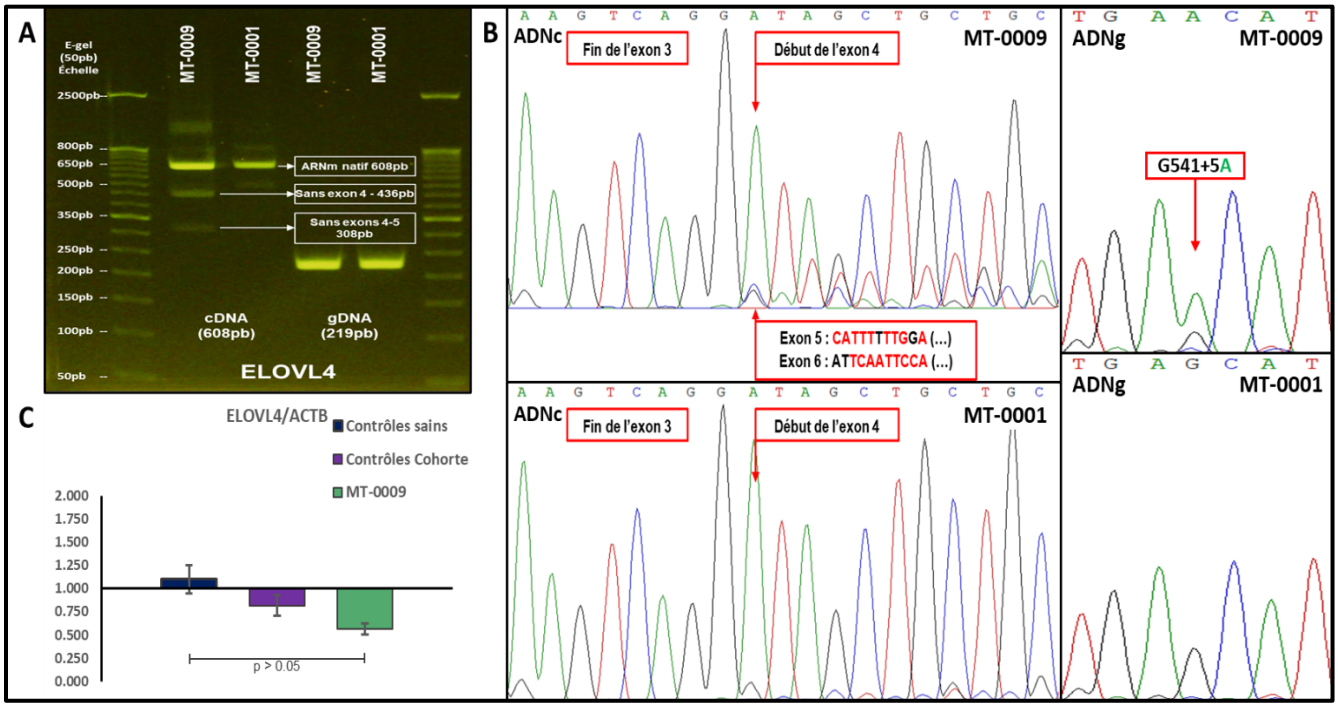

Figure 7. Validation expérimentale du variant *ELOVL4* chez MT-0009. Les régions flanquant c.541+5G>T dans l'ADN ainsi que dans l'ARNm sont amplifiés par PCR, puis les fragments résultants sont migrés sur E-gel 2% pour assurer la qualité des échantillons (**A**). Ceux-ci sont séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à un contrôle (**B**). Dans une plaque 384 puits, un triplicata contenant 15ng d'ADNc est incubé avec la sonde TaqMan de *ELOVL4*. La quantification du signal fluorescent par le système QuantStudio 6/7 est analysé en mode ΔΔCt relativement à l'expression de l'actine-β (**C**). Le groupe « sains » inclut MT-0001 et MT-0003 alors que « cohorte » inclut MT-0015 et MT-0011.

3.4.1.4 MT-0010

Suite à l'analyse des données WGS, trois variants avaient été sélectionnés pour l'étape de vérification expérimentale, soient un SNV faux-sens dans le gène *MARS* (c.G>A2210;p.R737Q) ainsi que deux variants causant potentiellement un épissage alternatif dans *PMPCB* (c.1154+5G>C) et *STAC2* (c.496-3T>G). Similairement aux candidats secondaires précédents, la vérification de *STAC2* a été mise de côté suite à l'absence de jonctions non canoniques dans les fragments migrés sur gel (non présenté) malgré la confirmation de la mutation intronique par Sanger. La confirmation du SNV non-synonyme hétérozygote de *MARS* (Figure 8A) est simplement moins intéressante que le candidat *PMPCB*. La migration de ce dernier sur E-gel (Figure 8A)

suggère la perte de l'exon 9 via AS étant donné le fragment distinct avec un poids moléculaire considérablement plus faible. Les résultats Sanger (Figure 8B) ont confirmé cette hypothèse, puisqu'il est possible d'observer une jonction des exons 8-10 qui est totalement absente du contrôle. La quantification par qPCR de *PMPCB* révèle quant à elle une diminution drastique des niveaux d'ARNm dans les PBMC du patient (Figure 8C). En effet, l'analyse préliminaire des niveaux d'expression suggère une baisse de plus de 76% comparativement aux contrôles hors-cohorte, qui demeure près de 52% lorsque la comparaison est faite avec deux contrôles de la cohorte EA. Ces observations demeurent toutefois, en fonction du faible nombre d'échantillons, significatives seulement avec le groupe sain selon l'ANOVA non-paramétrique à correction Tukey (p = 0.0003).

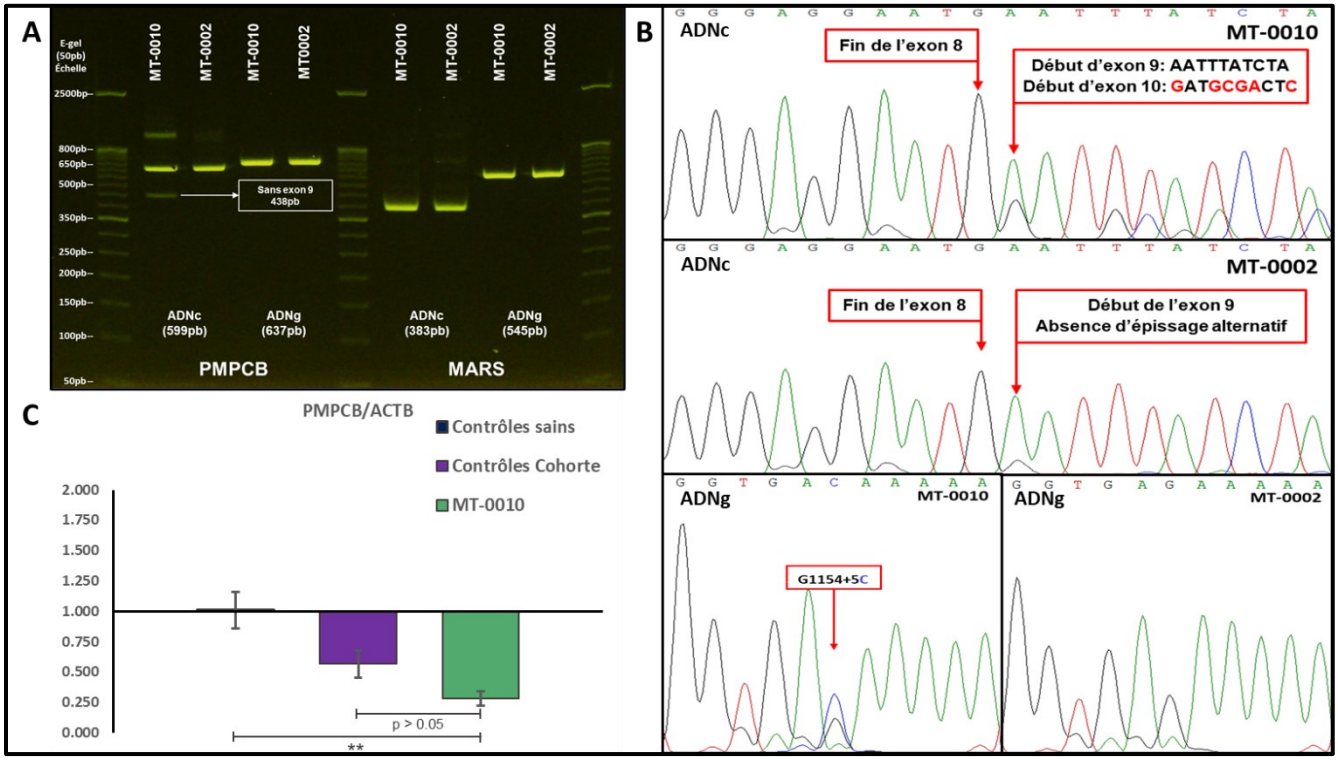

Figure 8. Validation expérimentale du variant *PMPCB* chez MT-0010. Les régions flanquant c.1154+5G>C dans l'ADN ainsi que l'ARNm sont amplifiés par PCR, puis les fragments résultants sont migrés sur E-gel 2% pour assurer la qualité des échantillons (**A**). Ceux-ci sont séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à un contrôle (**B**). Dans une plaque 384 puits, un triplicata contenant 15ng d'ADNc est incubé avec la sonde TaqMan de *PMPCB*. La quantification du signal fluorescent par le système QuantStudio 6/7 est analysé en mode ΔΔCt relativement à l'expression de l'actine-β (**C**). Le groupe « contrôles sains » inclut MT-0001 à MT-0003 alors que « contrôles cohorte » inclut MT-0015 et MT-0014.

3.4.1.5 MT-0011

Une seule mutation candidate a été sélectionnée pour une validation plus approfondie suite à l'analyse des variants chez MT-0011. Le gène *GABRP* présente un SNV faux-sens (c.445C>A;p.L149M) sur son 4^{e} exon. Malheureusement, comme la littérature le suggère, l'expression de ce gène dans les PBMC est nulle, ce qui a été confirmé par l'absence de lectures s'alignant à *GABRP* dans les données RNA-seq. La confirmation du variant n'a donc fonctionnée que sur l'ADNg, où l'amplification PCR permet la génération d'un fragment clair sur gel d'agarose 2% (Figure 9A). Aucun signal n'est détectable pour l'ADNc malgré une augmentation du nombre de cycles PCR. Bien que la bande ait moins migré comparativement à l'échelle moléculaire que ce qui est attendu, anomalie observée de façon consistante sur les gels d'agarose révélés par GelRed, celle-ci est la même pour le patient et le contrôle. De plus, le séquençage Sanger (Figure 9B) révèle bel et bien un fragment d'un peu moins de 600pb concordant avec la séquence génomique de *GABRP*, où le variant candidat hétérozygote est observable seulement chez MT-0011.

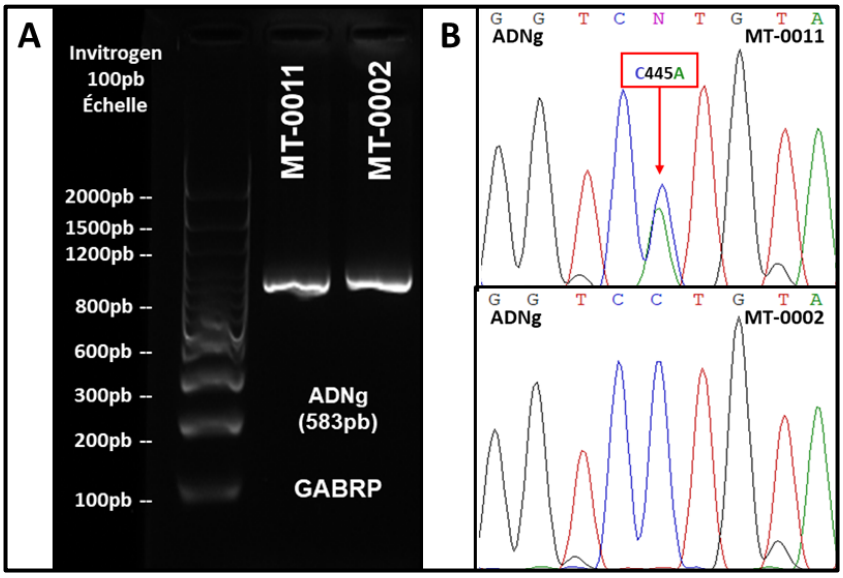

Figure 9. Vérification expérimentale du variant *GABRP* chez MT-0011. Les régions flanquant c.445C>A dans l'ADN génomique sont amplifiés par PCR. Les fragments résultants sont migrés parallèlement à l'échelle 100bp (Invitrogen™ #15628019) sur gel d'agarose traditionnelle 2%, révélé à l'aide d'agent intercalant GelRed (Biotium™ #41003), pour assurer la qualité des échantillons (**A**). Ceux-ci sont séquencés par technologie Sanger dans les deux sens, et la position du variant est visualisée par GeneStudio afin de comparer l'appel de base à un contrôle (**B**). Les captures d'écrans présentées correspondent au brin positif de l'ADNg.

3.4.1.6 MT-0012

Les deux candidats identifiés pour expliquer la pathologie de MT-0012 se trouvent sur le gène *SPG7*, soient un SNV non-sens (c.1861C>T;p.Q621X) ainsi qu'un faux-sens (c.2228T>C;p.I743T). En plus d'être présents dans les alignements WGS et RNA-seq (non présenté), les deux variants ont été amplifiés par PCR de l'ADNg ainsi que de l'ADNc. La migration sur E-gel (Figure 10A) permet d'observer des fragments clairs de taille attendue pour les deux régions génomiques contenant les variants. Pour l'ADNc, l'amplicon contenant simultanément les deux SNV possède également le bon poids moléculaire au niveau de la bande proéminente, malgré la présence de quelques fragments non spécifiques. Néanmoins, les résultats Sanger permettent d'observer un signal non ambigu pour la séquence d'ADNc, où il est possible de confirmer les deux variants hétérozygotes (Figure 10B). La même vérification a pu être faite pour l'ADNg, où les substitutions c.1861C>T ainsi que c.2228T>C sont présentes pour MT-0012, mais pas le contrôle (non présenté). En plus des candidats principaux, un SNV faux-sens identifié dans *ARHGAP4* (c.2294T>C;p.L765P) a été validé afin de possiblement expliquer un phénotype secondaire de diabète. Suite à l'amplification par PCR (Figure 10A), le variant a été confirmé par Sanger (non présenté). De plus, un autre SNV hétérozygote sur le gène *ATXN7L1* (c.40A>T;p.N14Y) a été confirmé (Figure 6A) afin d'appuyer la validation du gène chez MT-0007 et MT-0008, mais n'est pas candidat pour MT-0012.

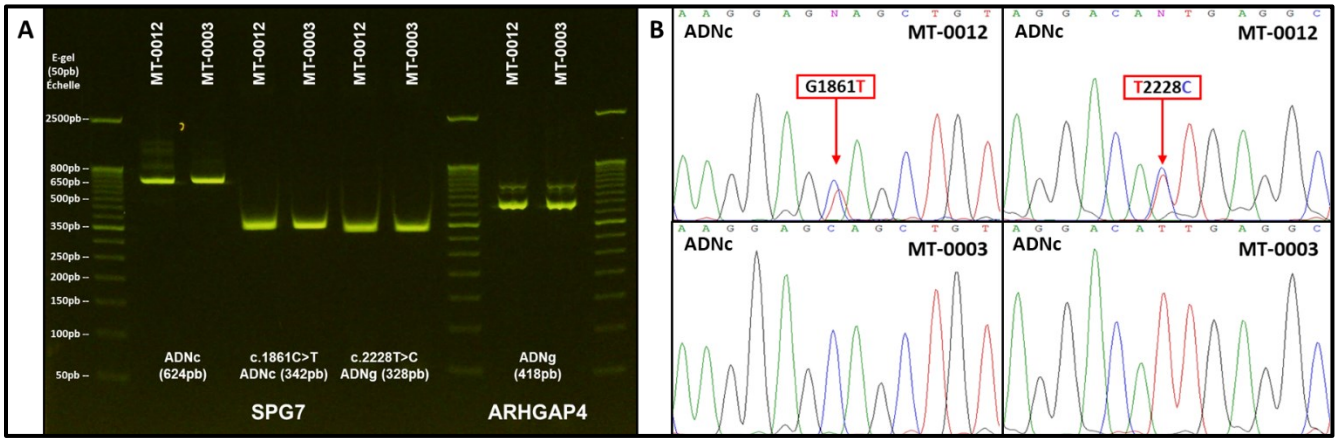

Figure 10. Vérification expérimentale des variants *SPG7* chez MT-0012. Les régions flanquant c.1861C>T et c.2228T>C dans l'ADN ainsi que dans l'ARNm sont amplifiés par PCR, tout comme c.2294T>C de *ARHGAP4*, puis les fragments résultants sont migrés sur E-gel pour assurer la qualité des échantillons (**A**). Ceux-ci sont séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à un contrôle, ici pour l'ADNc de *SPG7* (**B**).

3.4.1.7 MT-0013

Dans le but de trouver la cause de la pathologie de MT-0013, deux candidats ont été sélectionnés, soient une expansion *ATXN2* (CAG_{32}) ainsi qu'un SNV non-sens *ZFYVE26* (c.3022G>A;p.R1008X) étant potentiellement accompagné d'un AS prédit par rMATS. Malgré l'expression d'*ATXN2* dans le sang, confirmé par les données RNA-seq, seul l'ADNg a été amplifié pour confirmer la prédiction d'expansion poly-Q dans l'exon 1 faite par EH. Dès la migration sur gel (Figure 11A), il est possible de voir un dédoublement de la bande *ATXN2* chez le patient suggérant qu'un allèle a un poids moléculaire plus élevé que l'autre. De plus, la taille du fragment observée est supérieure à ce qui est attendue lorsqu'il n'y a que les 23 répétitions consensus (677pb). Les résultats Sanger du fragment contenant l'expansion permettent d'observer clairement des allèles de 32 ainsi que 31 répétitions CAG chez MT-0013, et cela pour les deux directions de séquençage (Figure 11C). Les données du contrôle ont malheureusement une trop faible qualité pour analyser cette information (non présenté). Du côté de *ZFYVE26*, l'amplification de la région contenant la troncation a également été un succès, alors que le potentiel épissage alternatif semble infirmé lors de la migration sur gel (Figure 11A). En effet, ces observations sont confirmées par la présence hétérozygote de c.3022G>A seulement chez MT-0013 (Figure 11B), alors que le profil de séquençage Sanger est quasi-identique chez le patient et le contrôle pour la région où de nouvelles jonctions d'épissage devraient être observées (non présenté). Malgré la confirmation de seulement un variant hétérozygote, la quantification des niveaux ARNm de *ZFYVE26* chez MT-0013 révèle une diminution considérable de ce dernier comparativement aux groupes contrôles (Figure 11D). En effet, l'analyse préliminaire des résultats suggère une baisse de plus de 58% comparativement aux contrôles hors-cohorte, ainsi qu'une diminution plus légère en comparaison avec deux contrôles de la cohorte EA (MT-0015 et MT-0011), à près de 45%. Selon l'ANOVA non-paramétrique avec correction Tukey, les changements sont significatifs ($p < 0.01$).

3.4.1.8 MT-0014

Suite à l'analyse des données du patient MT-0014, un seul gène candidat a été sélectionné pour une vérification expérimentale. Deux SNV hétérozygotes faux-sens sont retrouvés sur *CACNA1H* (c.4772G>A;p.R1591Q + c.2354A>T;p.K785M) et sont présent dans les alignements génomique autant que transcriptomique (non présenté). Une PCR a donc été réalisée pour l'ADNg ainsi que

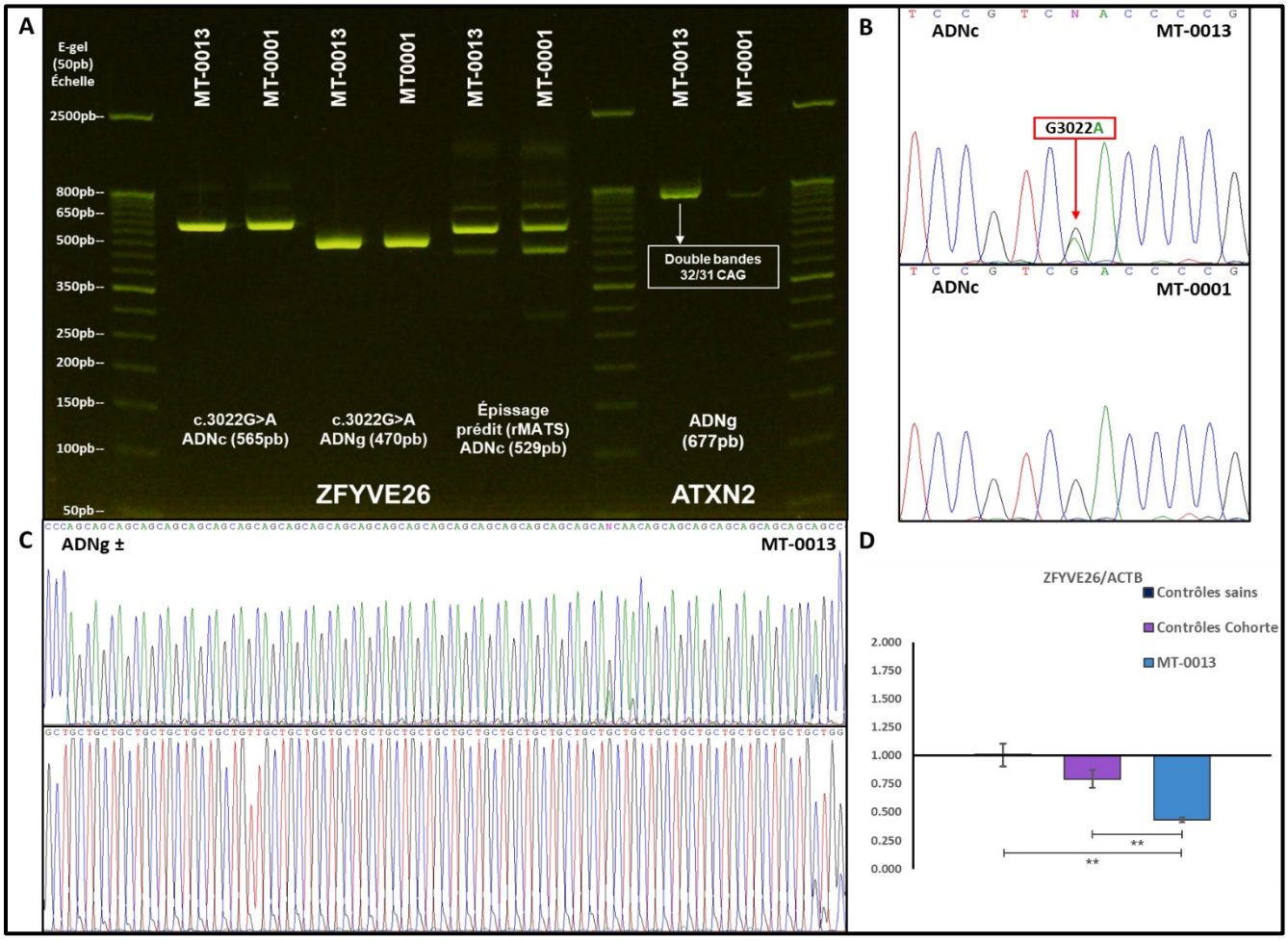

Figure 11. Validation expérimentale des variants *ATXN2* et *ZFYVE26* chez MT-0013. Les régions flanquant le STR CAG de *ATXN2* dans l'ADNg, ainsi que le variant c.3022G>A de *ZFYVE26* dans l'ADN et l'ARNm, sont amplifiés par PCR. Les fragments résultants sont migrés sur E-gel 2% pour assurer la qualité des échantillons, ainsi que pour une évaluation qualitative (**A**). Ils sont ensuite séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à un contrôle. La mutation *ZFYVE26* chez MT-0013 est comparée au contrôle MT-0001 (**B**) alors que seules les données du patient sont disponibles pour l'expansion *ATXN2* (**C**). Dans une plaque 384 puits, un triplicata contenant 15ng d'ADNc est incubé avec la sonde TaqMan de *ZFYVE26*. La quantification du signal fluorescent par le système QuantStudio 6/7 est analysé en mode ΔΔCt relativement à l'expression de l'actine-β (**C**). Le groupe « contrôles sains » inclut MT-0001 à MT-0003 alors que « contrôles cohorte » inclut MT-0015 et MT-0011.

pour l'ADNc (Figure 12A), cette dernière présente de nombreuses bandes inattendues suite à une amplification laborieuse. Le séquençage des fragments générés permet cependant de confirmer les deux variants de façon claire. Alors qu'un signal hétérozygote est attendu et observé pour c.4772G>A (Figure 12C, 12D), un débalancement allélique est vraisemblablement visible pour c.2354A>T (Figure 12B), où la mutation apparait de façon homozygote dans l'ADNc même si elle est clairement hétérozygote dans l'ADNg. Les variants sont bien sûr absents du contrôle MT-0002.

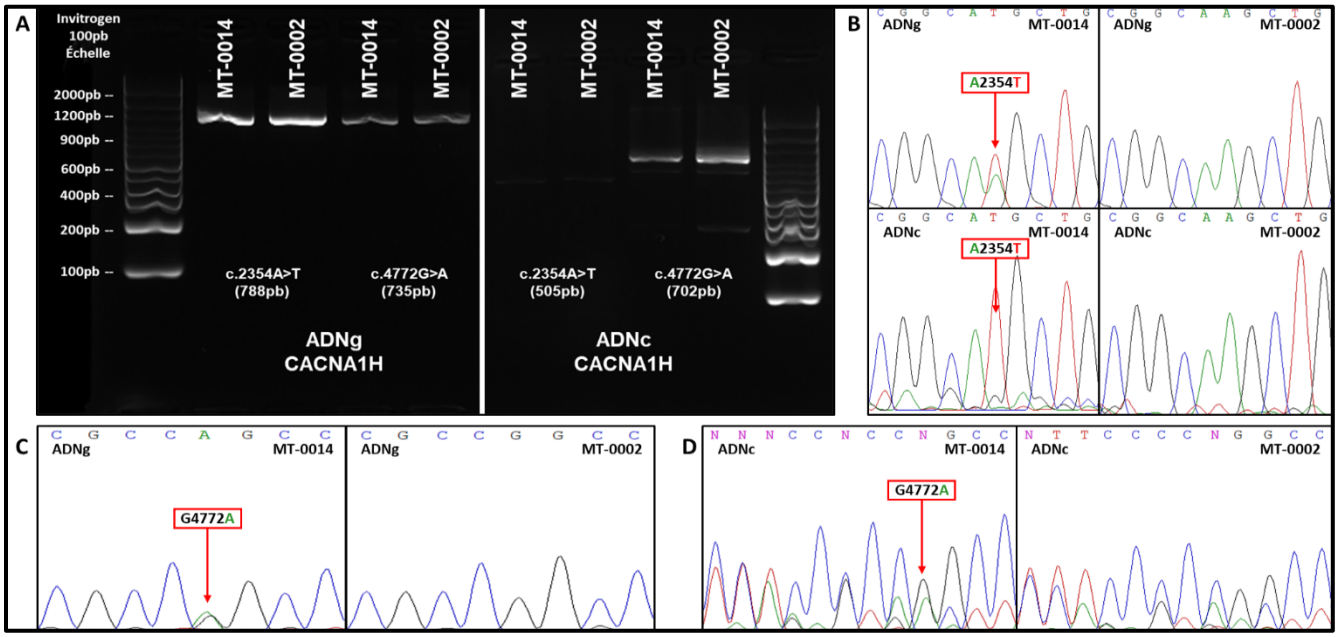

Figure 12. Vérification expérimentale des variants *CACNA1H* chez MT-0014. Les régions flanquant c.2354A>T et c.4772G>A dans l'ADN ainsi que dans l'ARNm sont amplifiés par PCR, puis les fragments résultants sont migrés sur gel d'agarose pour assurer la qualité des échantillons (**A**). Ceux-ci sont séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à un contrôle. On peut observer la mutation c.2354A>T dans les deux types d'acides nucléiques (**B**). Le variant c.4772G>A est visible dans l'ADNg (**C**), mais la présence de multiples transcrits le rend indiscernable dans l'ADNc (**D**). MT-0002 est un contrôle sain.

3.4.2 Validation du trio familial

Étant donné l'absence de RNA-seq pour confirmer la présence des variants candidats identifiés, la vérification par séquençage Sanger est d'autant plus importante, mais pas nécessairement possible pour une expansion de grande taille. Le traitement des données WGS est similaire à la cohorte EA, où l'analyse est non biaisée malgré l'hypothèse originale des cliniciens.

3.4.2.1 Variant *SPAST*

Suite à l'appel de variant via HaplotypeCaller de GATK, une SNV faux-sens dans le gène *SPAST* (c.1496G>A;pR499H) est rapidement devenu l'unique candidat pour expliquer la pathologie de la jeune patiente. La vérification de la mutation a été réalisée par l'amplification de la région d'intérêt dans l'ADNg (Figure 13A), suivi du séquençage Sanger du trio familial. Puisque le SNV hétérozygote n'est présent que chez la fille (Figure 13B), les résultats confirment une transmission *de novo* de la mutation. En plus d'être prédit comme pathogénique par plusieurs outils *in silico* (I-Mutant 3.0; MutationTaster), un alignement multiple de séquence permet de voir que la région codante est très conservée (Figure 13C), et le variant a déjà été rapporté comme « probablement pathogénique » dans la base de données ClinVar chez des patients atteints de HSP4.

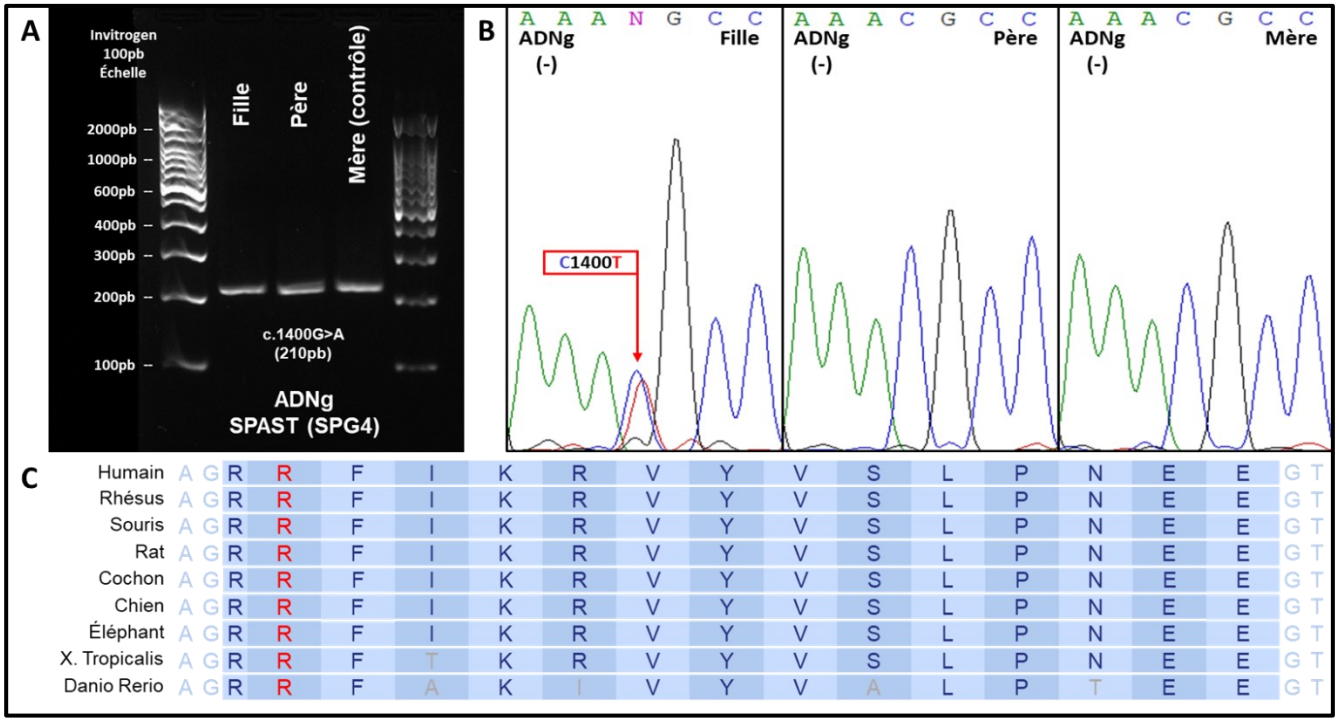

Figure 13. Vérification expérimentale du variant *SPAST* chez la fille française. Les régions flanquant c.1496G>A dans l'ADN génomique de la salive sont amplifiés par PCR, puis les fragments résultants sont migrés sur gel d'agarose 2% avec l'échelle moléculaire 100pb (Invitrogen) pour assurer la qualité des échantillons (**A**). Ceux-ci sont séquencés par technologie Sanger et les variants sont visualisés par GeneStudio afin de comparer l'appel de base à celui des parents (**B**). La conservation de la région codante est visualisable par alignement multiple de séquences provenant d'organismes variés, la plupart étant fréquemment utilisés en recherche (**C**).

3.4.2.2 Expansion *RFC1*

Pour expliquer la pathologie cérébelleuse du père, EH a été en mesure de prédire une expansion potentielle dans l'intron 2 de *RFC1*. En regardant l'alignement du WGS, une couverture largement réduite de la région est visible chez les présumés porteurs d'expansion AAAAG. Cette diminution semble proportionnelle au nombre d'allèles affectés, puisque la couverture quasi-nulle pour le père présumé homozygote est seulement réduite de moitié chez la fille hétérozygote, alors qu'elle est très bonne chez la mère en bonne santé (Figure 14A). Puisque la méthode Sanger ne permet pas un séquençage précis de longs fragments contenant des expansions, une évaluation qualitative du nombre de STR est effectuée par amplification PCR de la région d'intérêt, pour laquelle davantage de membres de la famille sont recrutés. Suite à la reconstruction du pedigree familial (Figure 14B), la migration des amplicons sur gel d'agarose permet d'observer clairement une expansion de plus de 500pb du côté paternel (Figure 14C). Puisque les répétitions sont de cinq nucléotides, les résultats suggèrent que le patient porte plus de 100 répétitions sur chaque allèle. Il est également possible de voir que l'amplicon du père subit la plus grande rétention, suggérant un nombre de répétitions supérieur chez le patient que chez les porteurs non atteints. Une qPCR a été tentée, mais les niveaux RFC1 étaient trop faibles dans la salive (non présenté).

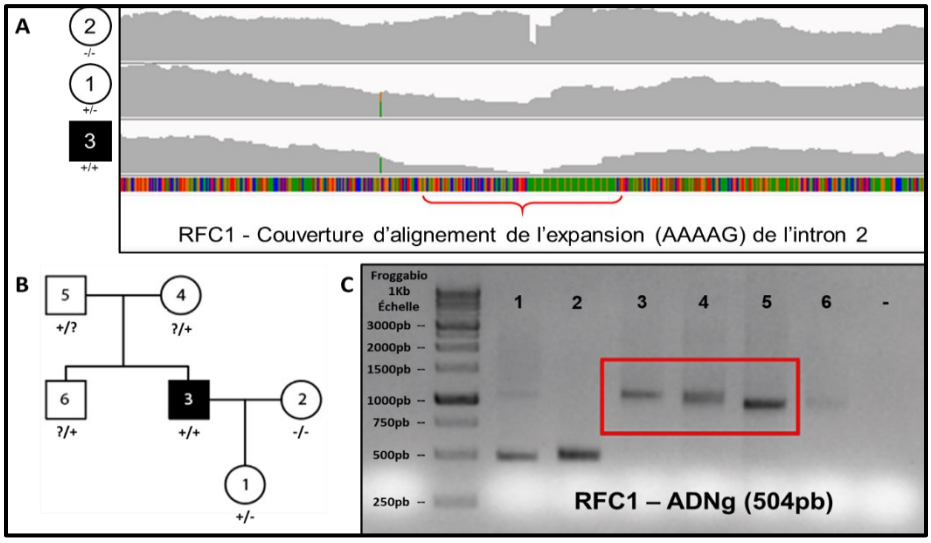

Figure 14. Vérification expérimentale de l'expansion *RFC1* chez le père. La couverture d'alignements WGS des trois échantillons est visualisée via IGV (**A**). Le pedigree des six membres de la famille, incluant un symbole pour indiquer si l'expansion est attendue sur les allèles (**B**). L'amplification PCR des régions flanquant le STR *RFC1* permet une évaluation qualitative de la taille des fragments suite à la migration sur gel d'agarose 1.5% (**C**).

4 - Discussion

4.1 Pertinence des PBMC chez les patients ataxiques

Puisqu'il n'est pas possible d'obtenir un échantillon provenant du tissu d'intérêt dans le cas des patients présentant des symptômes cérébelleux, il est important d'évaluer la pertinence du séquençage des PBMC dans le cadre d'études liées aux EA. Malgré le fait que les résultats obtenus par séquençage NovaSeq des PBMC ne correspondent pas parfaitement à ce qui est attendu dans les bases de données d'expression, l'observation de lectures alignées pour plus de 88% des gènes associés aux ataxies (Figure 4E) démontre la pertinence du tissu pour ce type de recherche. En effet, l'un des avantages majeurs des PBMC est que le prélèvement est beaucoup moins invasif que les autres options disponibles soient l'utilisation de fibroblastes ou encore de cellules musculaires obtenues par biopsies, qui n'assurent pas une plus grande correspondance au transcriptome cérébelleux. De plus, il est intéressant de noter qu'il y a un potentiel d'observer des changements d'expression statistiquement significatifs dans 84% des gènes (Figure 4F), information représentant fréquemment un premier indice d'impact fonctionnel d'une mutation. Cependant, l'approche n'est pas sans limitation, puisqu'il demeure difficile d'interpréter la pertinence d'un SNV détecté seulement via WGS, et pour lequel l'expression du gène affecté est nulle dans les PBMC. C'est le cas pour 4/15 gènes candidats de la cohorte EA, soient *SEC14L6* et *STAC2*, où les niveaux ARNm sont quasi-nuls, ainsi que *KCNA4* et *GABRP* qui ne sont pas exprimés. Pour ce dernier, candidat de MT-0011, cela complique l'obtention d'un diagnostic moléculaire sans prélèvement additionnel ou validation fonctionnelle exploitant un modèle animal/cellulaire. Un autre facteur à considérer est qu'une bonne partie des gènes précédemment identifiés dans la littérature (Tableau II) proviennent d'études initialement effectuées sur des échantillons sanguins, puisqu'il s'agit du standard pour ce type de recherche. Il est donc possible qu'une plus grande proportion des gènes dont l'association aux ataxies est toujours manquante aient une expression restreinte au cervelet, ce qui conduirait à un biais dans les statistiques présentées dans ce mémoire. Quoiqu'il en soit, les résultats suggèrent une forte pertinence de l'utilisation des PBMC pour l'obtention d'un profil transcriptomique des patients dans le cadre de ce projet.

4.2 Outils démontrant un grand potentiel pour la génomique clinique

La mise en place de pipeline d'analyse efficace pour les deux ensembles de données NGS, dans le but de pouvoir facilement identifier tous types de mutations, était l'un des objectifs secondaires du projet. En plus de l'implantation de nombreux outils bio-informatique standard, trois outils très récents ont été intégrés pour améliorer la prédiction d'épissage alternatif et permettre l'identification d'expansions de STR (66, 117, 119).

4.2.1 Outils de prédiction d'expansion nucléotidique

Les performances d'ExpansionHunter ainsi que STRetch ont été évaluées conjointement à l'aide d'un nombre considérable de contrôles positifs et négatifs. Ceux-ci permettent d'observer la spécificité ainsi que la sensibilité des prédictions effectuées par les outils (Figure 5). Cette dernière est très impressionnante pour EH, qui a détecté la quasi-totalité des expansions chez les contrôles positifs (32/34). Il est aussi important de constater que les échantillons pour lesquels l'appel est manquant sont simplement définis par l'outil comme ayant une trop faible couverture de la région pour une prédiction précise. Cependant, malgré le fait qu'il est effectivement possible d'observer une qualité non optimale de ces données, STRetch est en mesure d'appeler l'expansion *FXN* en question. Cela suggère une sensibilité imparfaite pour les deux outils, mais également qu'il pourrait être pertinent de les utiliser en tandem malgré une performance considérablement supérieur d'EH. En ce qui concerne la spécificité, définie dans ce cas-ci comme la capacité d'un algorithme à ne pas appeler un faux positif chez un contrôle négatif, la performance d'EH est très médiocre à première vue. C'est près de 40% des échantillons pour lesquels il y a au moins une prédiction d'expansion pathogénique n'étant pas réellement présente dans les données génomiques. Cependant, la majorité de ces faux positifs sont associés à seulement trois gènes appelés de façon récurrente comme portant plus de répétition que le seuil pathogénique défini dans la littérature soient *TCF4*, *NIPA1* ainsi que *PHOX2B*. Ces gènes sont plutôt reconnus comme facteurs de risques pour une dystrophie cornéenne à apparition tardive (146), la sclérose amyotrophique latérale (147), et une hypoventilation congénitale respectivement (148). Il est donc difficile de déterminer s'il s'agit effectivement de faux positifs, ou bien de découvertes fortuites chez des personnes présymptomatiques ou encore faiblement

affectés dû à une pénétrance variable de l'expansion. Puisque ces pathologies ne sont pas ou peu pertinente au type de patients soumis au pipeline d'analyse, et dans l'optique d'évaluer les performances sur des expansions dont la relation pathogénique est bien définie, l'ajustement de spécificité se traduit par une exclusion des appels de ces trois gènes dans le calcul du score.

Le faible nombre de contrôles négatifs pour STRetch s'explique quant à lui par le fait que 50 d'entre eux sont utilisés pour l'entraînement de l'outil, empêchant l'appel subséquent d'expansions. Étant donné les nombreux avantages d'EH en comparaison, il semblait peu pertinent d'évaluer la spécificité de STRetch de façon plus approfondie. Il est donc probable que le résultat présenté à cet effet sous-estime la précision de STRetch, qui possède tout de même un bon potentiel clinique. En plus de pouvoir appuyer une prédiction d'EH lors d'appels conjoints, l'outil offre une fonction intéressante calculant la fréquence d'observation de chaque STR dans un jeu de données. La capacité de discerner une augmentation très significative d'une répétition chez un patient, quelle qu'elle soit, est très pertinente à la recherche de nouvelles expansions, en plus de permettre la prédiction manuelle d'une expansion d'intérêt dans certains cas. Malgré le fait que seul EH soit en mesure de prédire une expansion pour le locus *RFC1* chez le trio familial, STRetch permet d'appuyer cette hypothèse en soulignant un nombre anormalement élevé du STR-AAAAG dans les WGS du père ainsi que de la fille (149). Cependant, EH offre plusieurs avantages considérables, tels qu'une plus grande facilité et rapidité d'exécution, une visualisation graphique des résultats suite à l'intégration de quelques scriptes, et une meilleure sensibilité.

Afin de discuter aussi des limitations, les outils tendent à fortement sous-estimer le nombre de STR d'une expansion, surtout STRetch, vraisemblablement dû au non-alignement des lectures trop courtes pour flanquer la répétition de nucléotides appariés. Aussi, il est fort probable que la sensibilité d'EH soient surestimées pour l'appel de STR pour lesquels peu de données d'entraînement bien définies sont disponibles, comme dans le cas de *TCF4* et *PHOX2B* où la rareté des mutations affecte les performances. En effet, puisque les contrôles positifs sont obtenus de bases de données publiques, l'évaluation de l'outil se limite à seulement quelques gènes. Il est donc essentiel de valider expérimentalement tous candidats identifiés par ce genre d'outils de prédiction. Toutefois, il sera intéressant d'observer l'optimisation des performances d'EH ainsi que l'ajout de nouveaux gènes cibles au courant des prochaines années, puisque de nouvelles

données d'expansions nucléotidiques s'ajoutent constamment à la littérature. Parallèlement, même s'il n'a pas été évalué lors du projet, EHdn semble offrir un potentiel similaire à STRetch pour la recherche plus exploratoire de causes pathogéniques, surtout dans les cas où une expansion commune est supposée chez plusieurs patients (118). L'interprétation des résultats est plus difficile, mais l'optimisation fréquente de l'outil pourrait rapidement mener à son implémentation future en recherche clinique. Quoiqu'il en soit, les résultats présentés suggèrent fortement l'intérêt d'intégrer conjointement EH et STRetch lors d'analyses clinique, afin de permettre une détection rapide et précise de candidats tels qu'*ATXN2* chez MT-0013.

4.2.2 Outil de prédiction d'événements d'épissage alternatif

L'évaluation appliquée des performances des outils de prédiction d'AS est plus complexe qu'une simple évaluation comparative de détection d'événement, et plusieurs articles se penchent déjà sur la sensibilité et la précision de SpliceAI (125, 150, 151). Pour ces raisons, l'interprétation de la valeur de l'outil pour un pipeline de génomique clinique se limite ici à une simple analyse qualitative des résultats obtenus avec les données de patients des deux projets. Plusieurs informations peuvent être extraites des statistiques présentées (Tableau VI), entre autre le clair avantage de l'utilisation sur WGS puisque la majorité des mutations captées par séquençage transcriptomique le sont également par séquençage génomique. En effet, puisque l'outil ne base pas ses appels sur la détection d'événements réels (count-based), mais plutôt sur le potentiel d'un changement nucléotidique à perturber des sites donneurs et accepteurs adjacents, il est l'un des seuls outils du genre à ne pas se limiter aux régions exprimées. Il est également intéressant d'observer que près de 85% des prédictions concernent des mutations non codantes malgré le fait qu'une bonne partie de la littérature portant sur les variants d'épissages pathogéniques implique les régions codantes ainsi que les jonctions d'épissages (76). Cela signifie que malgré le fait que le réseau neural de SpliceAI apprend principalement à partir de données codantes, le biais de priorisation attendu est vraisemblablement mitigé par la profondeur de l'algorithme (66). Un autre aspect intéressant de SpliceAI est qu'il donne simultanément le score des quatre types d'événements possibles, en plus de la position du site alternatif potentiel, favorisant une interprétation simple et organisée des résultats obtenus pour chaque échantillon. À cet effet, l'utilisation de différents seuils d'appel permet une grande variété d'applications, où un filtre

indulgent (0.2) permet la capture de nombreux variants dans des gènes d'intérêt ciblés, alors qu'un filtre plus strict (0.85) mène à l'obtention rapide des événements d'AS fréquemment réels (confirmation via données RNA-seq lorsqu'exprimés). Avec un peu moins d'une quarantaine de candidats par échantillon de WGS lorsque le seuil de 0.85 est utilisé, l'intégration de l'outil à un pipeline clinique n'implique qu'une faible charge d'analyse supplémentaire. Toutefois, l'exécution des prédictions SpliceAI nécessite une quantité considérable de ressources computationnelles ainsi que plusieurs jours pour des données de séquençage génomique. Alors que ne pas se baser sur la détection d'événements comptés (count-based) est bénéfique en termes d'applicabilité à tous les types de données NGS, cela représente aussi une limitation majeure de la méthode, qui dépend donc fortement des annotations de transcrits disponibles, et des connaissances sur l'épissage canonique d'un gène. En plus de naturellement générer de nombreux faux positifs, SpliceAI ne peut considérer l'existence d'isoformes non canoniques qui seraient dus à une annotation incomplète dans la littérature (152).

En ce qui concerne les données des patients, SpliceAI a correctement prédit l'effet fonctionnel des variants *ELOVL4* et *PMPCB* sur l'épissage de leur ARNm, alors que rMATS n'a détecté que ce dernier dû à une faible couverture d'*ELOVL4*. À l'opposé, alors que rMATS appuyait l'hypothèse d'un épissage alternatif de *ZFYVE26* chez MT-0013, qui s'est avéré être incorrecte, SpliceAI n'a pas fait cet appel. De plus, plusieurs candidats secondaires ont été identifiés chez certains patients (non présentés), mais non pas été confirmés par les données RNA-seq vu leur expression trop faible, et pourraient devenir pertinents si les candidats principaux sont invalides. L'évaluation générale de l'outil est donc en ligne avec la littérature (151), avec des résultats qui suggèrent que SpliceAI a un énorme potentiel en recherche clinique. Malgré l'avantage flagrant d'utilisation sur les données génomiques, l'intégration peut être bénéfique à tout pipeline d'analyse NGS. Aussi, puisque l'algorithme est basé sur l'apprentissage profond par intelligence artificielle, il est fort probable que l'outil continue de s'améliorer au travers des futures itérations exploitant une littérature plus complète. Finalement, le potentiel de SpliceAI s'étend au-delà de la recherche de mutations pathogénique : la découverte de nouveaux variants cryptiques ou non codants ayant la capacité d'affecter le transcriptome permet d'améliorer la compréhension des facteurs régulant l'épissage, pouvant subséquemment mener à l'interprétation de nombreux VUS.

4.3 Candidats finaux de la cohorte EA

4.3.1 SNV faux-sens dans *ATXN7L1* & *SEC14L6*

Quoique les deux patients du duo père-fils (MT-0007 et MT-0008) présentaient certains variants distincts, et qu'il est toujours possible qu'une mutation *de novo* soit à l'origine d'une pathologie différente chez le fils, l'optique d'une cause commune s'avère ici plus pertinente. La mutation *ATXN7L1* (c.2059G>A;p.A687T) représente un bon candidat pour plusieurs raisons, la plus évidente était la forte homologie du gène avec *ATXN7*, cause bien définie de SCA de type 7 dont les phénotypes typiques concordent très bien avec les deux patients (157). Alors que la majorité des cas SCA7 sont liés à une expansion poly-Q, certains SNV de *ATXN7* classifiés comme VUS ont été observés chez des patients atteints de cette maladie, suggérant qu'une mutation non-synonyme pourrait possiblement être responsable des phénotypes de MT-0007 et MT-0008 (Tableau III). Puisqu'*ATXN7L1* partage le domaine fonctionnel « SCA7 » d'*ATXN7*, il est possible qu'il participe également à la fonction de stabilisation des microtubules (34). Le patron d'expression des deux gènes est très similaire (134), tout comme leurs réseaux d'interaction protéique (136), appuyant davantage l'hypothèse d'une fonction redondante. Au niveau du variant lui-même, une faible fréquence allélique ($6.56x10^{-5}$) qui s'étend sur le gène, lui conférant une faible tolérance (pLI = 1) aux mutations (106), une forte conservation des acides aminés de l'exon à travers de nombreuses espèces (non présenté), ainsi qu'une prédiction CADD appréciable (score = 23.8) supportent le potentiel pathogénique du changement d'acide aminé. Aussi, il est intéressant de noter qu'alors que c.2059G>A est présent dans quatre des cinq isoformes connus d'*ATXN7L1*, le VUS c.40A>T identifié chez MT-0012 est en plus absent du transcrit NM_001318229. La possibilité que ce dernier ait une importance spécifique à certains tissus est la principale raison derrière son inclusion à la validation du candidat. Celle-ci inclut la confirmation définitive du variant dans l'ADN et l'ARNm par séquençage Sanger (Figure 6A, 6C) ainsi qu'une évaluation préliminaire de l'effet fonctionnel du variant sur l'expression d'*ATXN7L1* (Figure 6B). Quoique l'analyse des résultats démontre une diminution considérable comparativement aux contrôles sains, ceux-ci présentent des niveaux d'ARNm étant différents à ceux du contrôle de la cohorte (MT-0015) de façon consistante. Par souci de pertinence, c'est donc ce dernier qui est

priorisé pour déterminer si un effet sur l'expression d'un gène est présent. Cela fait en sorte que la diminution d'*ATXN7L1* chez les deux patients est statistiquement non significative, ce qui est en ligne avec l'analyse d'expression différentielle effectuée sur les données RNA-seq où la diminution moyenne était d'environ 10% (Valeur Log2 = -0.138). Cependant, il serait important de refaire l'expérience avec davantage de contrôles, afin de confirmer les résultats de façon plus solide. L'absence de changement du niveau d'expression ne signifie nécessairement pas que le variant n'impacte fonctionnellement *ATXN7L1*, mais il est certain qu'une validation plus approfondie sera essentielle à l'association du gène aux phénotypes des patients (12).

Le SNV faux-sens identifié dans l'exon 11 de *SEC14L6* (c.1021C>A;p.R341S) est également intéressant malgré l'absence de lien direct avec une pathologie ataxique connue. L'inférence d'une possible implication dans l'AVED, normalement associée à une mutation du gène *TTPA* (25), est principalement due au fait qu'il y a une forte similitude de séquence entre cette protéine liant l'α-tocophérol et les membres de la famille *SEC14*. Le réseau d'interaction protéique projeté possède également de nombreuses ressemblances, entre autres la proximité des voies d'hydrolyse de phospholipides par *PLD2* ou encore *SACM1L*, qui interagissent directement avec *SEC14L6 (136)*. Le patron d'expression diffère cependant de *TTPA*, gène étant majoritairement concentré dans le foie comparativement à *SEC14L6* qui est principalement exprimé dans le cerveau et la glande thyroïde (132). Néanmoins, il ne serait pas surprenant que l'altération d'une fonction de liaison des lipides (158), prédite chez *SEC14L6* dû à son domaine *SEC14*, soit en mesure d'induire des phénotypes neurologiques. En effet, le métabolisme des lipides joue plusieurs rôles importants dans les neurones (159), et son implication dans certains désordres neurologiques est déjà bien établie (58). Encore une fois, l'AF du variant est faible ($1.19x10^{-5}$) et la prédiction d'altération fonctionnelle par CADD est élevée (score = 24), mais la tolérance de mutations pour *SEC14L6* est généralement bien plus grande que pour *ATXN7L1* (pLI = 0). Il est aussi important de noter que la mutation n'est présente que sur deux des trois isoformes connus du gène, et qu'elle ne se retrouve pas dans le domaine *SEC14* mais bien dans une partie de la séquence associée aux dynamiques du transport vers le Golgi (GOLD), qui peut tout de même être essentielle au fonctionnement de la protéine (160). Pour ces raisons, la vérification du candidat, second seulement à *ATXN7L1*, a été réalisée sur les deux types d'acides nucléiques (Figure 6D, 6F)

malgré la faible expression des transcrits dans le sang. Le séquençage a permis d'observer que l'exon d'intérêt fait partie des isoformes mineurs de *SEC14L6*, du moins dans le sang où la quantification d'expression est peu conclusive dû aux très faibles niveaux d'ARNm (Figure 6E). Malgré une potentielle diminution importante chez les patients, l'utilisation d'un tissu ayant de bons niveaux d'expression de *SEC14L6* est obligatoire à l'interprétation de l'effet fonctionnel du variant sur le transcriptome du gène, afin de mitiger la déviation due aux manipulations. Pour ces raisons, l'accent des validations futures sera sur le variant *ATXN7L1*, malgré l'intérêt d'évaluer l'impact du variant dans un tissu plus pertinent à *SEC14L6*, tel que les fibroblastes. À noter, il est possible que l'exon 11 soit non-essentiel aux fonctions du gène, ce qui invaliderait quasi-assurément l'hypothèse d'un effet pathologique similaire à l'AVED chez les patients MT-0007 et MT-0008.

Comme mentionné dans les résultats, *VPS13C* a été investigué chez MT-0008, mais plus spécifiquement dans le but d'expliquer les phénotypes différentiels du patient (Tableau III). Le gène en question est reconnu comme causant une forme AR précoce de Parkinson (type 23), associée non seulement à des tremblements, mais également certains symptômes atypiques qui distinguent MT-0008 tels qu'une légère spasticité, une hyperréflexie ainsi qu'une incontinence urinaire (153). Les variants c.11014G>T et c.4165+1192A>C ont donc été évalués expérimentalement (non présenté), mais ne sont que secondaires à la validation clinique, ayant plutôt un potentiel de modulateur génétique (49).

Dans le cas de la mutation *KCNA4*, c'est plutôt l'absence d'expression du gène dans les PBMC qui fait en sorte que les variants communs demeurent prioritaires pour l'obtention éventuelle d'un diagnostic. Puisque *KCNA4* est un bon paralogue de *KCNA1*, et participe à la formation de l'hétérotétramère fonctionnel agissant comme canal potassique, le gène est associé aux ataxies épisodiques par inférence (154, 155). Alors que cela en fait un excellent locus candidat pour expliquer les phénotypes du père, la confirmation du variant intronique c.-783+5G>T dans les données WGS ainsi que l'ADNg (non présenté) ne suffit pas à l'appui d'un effet fonctionnel sur le transcriptome de *KCNA4*. Étant donné que l'expression du gène est quasi-exclusive au cerveau, la transdifférenciation de cellules en neurones serait nécessaire à la validation de l'implication du variant dans la pathologie de MT-0008 (156). Comme l'hypothèse provient seulement d'une

prédiction de SpliceAI, où le variant induit potentiellement un gain de donneur (score = 0.28) pouvant causer une rétention de région UTR par exemple, la vérification d'AS serait intéressante, mais est secondaire aux mutations communes à MT-0007. Avec un score aussi faible, les chances d'obtenir une validation négative sont considérables.

4.3.2 Épissage alternatif d'*ELOVL4*

L'intérêt pour le variant identifié dans *ELOVL4* (c.541+5G>A) est bien plus évident, même s'il se retrouve dans une région non codante. La mutation du gène en question est déjà une cause établie de SCA de type 34 (161), une forme AD dont les symptômes typiques d'ataxie lentement progressive correspondent quasi-parfaitement à ceux de MT-0009, en plus d'inclure certaines caractéristiques du patient telles qu'une hyperréflexie, une atrophie cérébelleuse, plusieurs phénotypes visuels, ainsi que des tremblements légers (Tableau III). *ELOVL4* est un gène impliqué dans la biosynthèse des lipides dont l'expression est restreinte à la rétine, au cerveau, au thymus, ainsi qu'aux cellules de la peau (132). Tel que mentionné, le métabolisme des lipides semble essentiel au bon fonctionnement du système nerveux (159), ce qui explique pourquoi une dysfonction d'*ELOVL4* est en mesure d'induire une dégénérescence neuronale (161). La prédiction de pathogénicité est tout de même élevée (score = 19.41), d'autant plus si on considère la nature intronique du variant, pour lequel ces scores ont tendance à être plus faible. La fréquence allélique à cette position chromosomique est complètement nulle, et la tolérance mutationnelle d'*ELOVL4* est très faible (pLI = 0.83).

Un fait intéressant est que c.541+5G>A n'est détecté par HaplotypeCaller et SpliceAI (AL = 0.43; DL = 0.65) que pour le WGS, puisque la couverture de la position est trop faible dans le RNA-seq. Ce dernier permet cependant tout de même de voir le variant (3/3 lectures), en plus de permettre d'observer des jonctions non canoniques d'exons 3-6, appuyant la prédiction de perte de donneur (exon 4) et accepteur (exon 5) par SpliceAI, ce qui cause un double saut d'exon. Le déséquilibre transcriptomique (Figure 7A), où deux fragments additionnels sont visibles sous la bande canonique de 608pb, ainsi que les données Sanger démontrant les jonctions 3-5 et 3-6 (Figure 7B) confirment les deux événements distincts de perte d'exon 4 (436pb) ainsi que de perte des exons 4 et 5 (308pb). C'est donc dire que le RNA-seq à lui seul ne permet pas d'identifier le variant, mais

que le WGS n'offre que de bons scores prédictifs, où l'absence de preuves sur la possibilité d'AS préalablement à la validation expérimentale pourrait mener à une classification du VUS comme candidat de second plan. Ce type d'observation appuie l'hypothèse où la combinaison des deux technologies est bénéfique à l'investigation des cas complexes.

La diminution d'à peu près 50% des transcrits *ELOVL4* chez le patient (Figure 7C) corrèle avec l'hypothèse de dégradation médiée par ubiquitination (NMD) de l'allèle muté. Étant donné le fait que SCA34 est une pathologie AD, et que les exons 4-5 sont vraisemblablement essentiels à la fonction de la protéine puisqu'ils encodent plusieurs domaines transmembranaires, il est plus que probable que la perte fonctionnelle complète d'un allèle soit responsable des phénotypes ataxiques de MT-0009. À cet effet, plusieurs mutations touchant l'exon 4 sont déjà rapportées dans la littérature comme causatif d'ataxie (162). Malgré le fait qu'une simple confirmation du variant par un laboratoire clinique devrait suffire au diagnostic moléculaire du patient (12), il serait potentiellement intéressant d'investiguer si la nature unique du variant présenté ici est à l'origine de l'aspect épisodique des phénotypes de MT-0009. c.541+5G>A est le premier variant d'épissage affectant l'exon 4 d'*ELOVL4*, et comparativement à un SNV faux-sens, ceux-ci sont susceptible de causer une insuffisance allélique plutôt qu'un fonctionnement anormal de la protéine (76). L'hypothèse serait donc que l'allèle sain est en mesure de rescaper partiellement le phénotype. Bien évidemment, malgré le fait que *KCNAB3* ait été validé expérimentalement (non présenté) dû à sa fonction de canal potassique (163), les résultats obtenus pour *ELOVL4* ont mis fin à l'investigation de ce variant, du moins pour le moment.

4.3.3 Épissage alternatif de *PMPCB*

Similairement à MT-0009, le variant intronique de *PMPCB* (c.1154+5G>C) a été identifié dans les données de WGS par HaplotypeCaller ainsi que SpliceAI (AL = 0.74; DL = 0.98) comme causant potentiellement la perte de l'exon 9. La présence de jonction des exons 8-10 dans l'alignement RNA-seq permet d'appuyer cette hypothèse (non présenté). Plusieurs éléments supportent la candidature de *PMPCB* pour expliquer la pathologie du patient, dont un lien direct avec le syndrome de dysfonctions mitochondriales multiples de type 6 (169). Le patient présente un chevauchement considérable avec les phénotypes associés à cette maladie (Tableau III), incluant

des symptômes atypiques de spasticité ainsi que de perte auditive. De plus, *PMPCA*, son partenaire fonctionnel et paralogue, est une cause établie d'ataxie de SCAR de type 2 où certains phénotypes concordent également avec MT-0010 (170). Cependant, la présentation clinique semble considérablement moins sévère et plus tardive que les patients typiques de ces deux désordres, qui occasionnent généralement des anormalités développementales. À cet effet, il est important de noter que ceux-ci sont habituellement AR, signifiant que la validation d'une forme AD pourrait soutenir une dysfonction moléculaire partielle ainsi qu'une pénétrance variable des phénotypes (171). Vu la nature faux-sens de la majorité des variants AR rapportés (43), l'identification d'un AS majeur pourrait suffire à l'apparition de phénotypes (76). *PMPCB* est l'une des deux sous-unités catalytiques d'une protéase mitochondriale essentielle au clivage d'autres précurseurs de protéines mitochondriales telles que *PINK1* et *FXN* (172), et son expression est ubiquitaire (132). Il est alors cohérent qu'une perturbation importante de multiples fonctions de la mitochondrie engendre une atteinte neurologique majeure avec phénotypes d'ataxies (56, 57).

En ce qui concerne le variant lui-même, la prédiction CADD est bonne (score = 16.7), l'AF est faible ($4.91x10^{-5}$), et le score de tolérance de *PMPCB* est mauvais (pLI = 0). Dès la migration sur gel (Figure 8A), il est possible de confirmer l'événement d'AS dans l'ADNc, où un fragment sans exon 9 (438pb) est visible sous la bande attendue (599pb). Alors que le séquençage Sanger confirme le variant hétérozygote dans l'ADNg, la jonction d'exons 8-10 est visible dans l'ADNc (Figure 8B). Étant donné la présence de l'exon 9 dans la quasi-totalité des transcrits canoniques selon la base de données GTEx (132), il est probable qu'un effet fonctionnel découle de cette altération transcriptomique. À cet effet, la quantification des niveaux *PMPCB* dans les PBMC par qPCR permet encore une fois d'observer une baisse d'expression de plus de 50%, suggérant que l'allèle muté voit son ARNm être dégradé par NMD. Bien sûr, une validation plus approfondie est nécessaire afin de prouver l'implication du variant *PMPCB* dans la pathologie ataxique de MT-0010 (12), mais les résultats préliminaires corrèlent avec l'hypothèse d'une dysfonction mitochondriale pathogénique. Il serait intéressant de confirmer l'impact sur la protéine par combinaison d'un immunobuvardage traditionnel avec une évaluation de l'efficacité de la protéase mitochondriale par observation quantitative du clivage fonctionnel sur l'une de ses nombreuses cibles (170)

Suite à l'analyse des données de séquençage, trois candidats avaient été sélectionnés pour l'étape de validation fonctionnelle. Un variant dans une région d'épissage de *STAC2* (c.496-3T>G) engendrait une forte prédiction de déplacement du site accepteur (AG = 1.00; AL = 0.98), ce qui a le potentiel d'induire un changement de cadre de lecture. Puisque le gène a une fonction connue de modulation du canal Cav2.1 (*CACNA1*) *(164)*, et une association indirecte à une pathologie épisodique musculaire (165), un lien avec les ataxies épisodiques était possible. Malgré le fait que l'expression de *STAC2* était non-nulle, quoique minimale, la vérification par PCR ne permet que de confirmer l'existence du variant dans l'ADNg (non présenté). Il est difficile d'évaluer un impact transcriptomique à partir des PBMC, puisque *STAC2* n'est exprimé à des niveaux appréciables que dans le cerveau, la peau, ainsi que les tissus sexuels. Cela signifie que l'absence d'AS suite à l'amplification des quelques transcrits d'ARNm disponibles (non présenté) ne suffit pas à conclure que le variant n'impacte pas l'épissage de *STAC2*, mais que la validation du variant est secondaire aux autres candidats de MT-0010.

La mutation faux-sens identifiée dans *MARS* (c.2210G>A;p.R737Q) est quant à elle intéressante compte tenu de l'implication directe du gène dans la neuropathie de Charcot-Marie-Tooth (166), et indirecte dans une ataxie spastique AR via une homologie à *MARS2* (167). Le gène joue un rôle critique dans la biosynthèse des protéines puisqu'il permet de charger les ARNt avec leurs acides aminés respectifs (168). Il n'est donc pas surprenant que son expression soit ubiquitaire, quoique maximal dans le cerveau (132). La prédiction CADD est très élevée (score = 34) et l'AF du variant est nulle, quoique six porteurs de R373W sont répertoriés sur gnomAD (106). Cependant, il est prédit que *MARS* tolère bien les mutations (pLI = 0). La vérification PCR sur l'ADNg ainsi que l'ADNc s'est bien déroulée (non présenté), et le gène est toujours un bon candidat pour expliquer la pathologie de MT-0010, mais l'intérêt pour le SNV faux-sens est légèrement moindre puisqu'il est complexe de démontrer un effet fonctionnel causé par le remplacement d'un seul acide aminé.

4.3.4 VUS dans l'ADNg pour *GABRP*

L'analyse des données WGS n'a généré qu'un candidat intéressant pour MT-0011, soit un SNV faux-sens dans l'exon 5 du gène *GABRP* (c.445C>A;p.L149M). Il s'agit de la sous-unité π du récepteur d'acide γ-aminobutyrique (GABA), neurotransmetteur inhibiteur principal du système

nerveux central (173). Un aspect cryptique du gène est qu'alors que son expression d'ARNm est faible dans le cerveau, les niveaux protéiques observés sont très élevés, et ce surtout au niveau du cervelet (132). L'attrait principal du candidat est son association par inférence aux EA de type 4 dû à des études génomiques d'association (174), mais aucun lien direct avec une pathologie n'est établi. Puisque la mutation se retrouve dans l'exon 5, qui fait partie du domaine extracellulaire de la protéine, l'hypothèse serait qu'une altération des interactions canoniques de la protéine altère la signalisation synaptique de *GABRP* (175). Le fait que l'exon 5 ne soit pas exprimé dans tous les isoformes du gène pourrait participer à la spécificité tissulaire du désordre. Néanmoins, une investigation extensive serait nécessaire afin de prouver une telle association génotype-phénotype. Le SNV a un bon potentiel pathogénique selon CADD (score = 25.7), et l'AF est faible ($1.19x10^{-5}$), mais le gène est prédit comme ayant une bonne tolérance aux mutations (pLI = 0). Malheureusement, les données RNA-seq confirment que *GABRP* n'est pas exprimé dans les PBMC, conduisant à une vérification expérimentale restreinte à l'ADNg. L'amplification par PCR (Figure 9A) suivie du séquençage Sanger (Figure 9B) a bien confirmé la présence hétérozygote du variant faux-sens, mais la validation n'a pas été poussée plus loin pour le moment. La recherche de candidats alternatifs grâce aux nombreux outils de prédictions est toujours en cours.

4.3.5 Double SNV pathogénique dans *SPG7*

Pour MT-0012, l'analyse des résultats de HaplotypeCaller a permis d'identifier rapidement la cause moléculaire du désordre génétique du patient. Deux SNV hétérozygotes sont présents dans le gène *SPG7*, soit l'induction d'un codon d'arrêt (c.1861C>T;pQ621X) et une mutation faux-sens (c.2228T>C;p.I743T). Ce qui permet de promptement conclure qu'il s'agit du gène causatif est que *SPG7* est une cause génétique bien connue de HSP de type 7, forme généralement AR, mais où des cas AD sont également répertoriés (171). Quoi qu'il en soit, les deux variants possèdent distinctement un grand potentiel pathogénique puisque les troncations protéiques ont fréquemment un impact fonctionnel important (176), et que p.I743T a précédemment été identifié chez plusieurs patients atteints de HSP7 (177). De plus, il s'agit d'un gène candidat idéal pour MT-0012, puisque les phénotypes caractériels de la maladie concordent non seulement parfaitement avec le patient (Tableau III), mais *SPG7* est une cause récurrente d'ataxies spastiques dans la population canadienne (178). Le variant non-sens est prédit comme fortement

pathogénique par CADD (score = 39) et est décrit pour la première fois dans ce projet, appuyant son potentiel délétère. Le SNV faux-sens obtient également un score de pathogénicité élevé (score = 25.4) et est associé à seulement 14 porteurs sur gnomAD ($4.95x10^{-5}$), correspondant seulement au double du nombre patient HSP7 avec cette mutation dans ClinVar (43, 106). Les candidats ont bien sûr été confirmés au niveau de l'ADN et de l'ARN (Figure 10).

Comme il s'agit de deux variants codants causant un changement de séquence protéique, et présent dans un gène bien défini de HSP avec phénotypes spastiques (Tableau II), le diagnostic aurait vraisemblablement dû être réalisé à l'aide d'un test par panel contrairement aux autres patients (Tableau VII). Cela étant dit, il est possible que le panel ayant été sélectionné pour évaluer la cause génétique de MT-0012 n'incluait pas *SPG7*, puisque malgré la forte composante ataxique de HSP7, la maladie demeure principalement classée comme une paraplégie spastique. En dépit de son coût plus élevé, cela souligne un énorme avantage du NGS, qui n'est pas restreint à un nombre prédéterminé de candidats et diminue les risques de rater un diagnostic moléculaire pouvant bénéficier la prise en charge du patient (10). Il y a également réduction du fardeau médical concernant le choix du panel génique optimal par le clinicien, qui repose sur une prédiction précise de la pathologie neurologique parmi d'innombrables formes cliniques ayant des présentations similaires.

Dans cette même optique, un SNV faux-sens a été identifié dans *ARHGAP4* (c.2294T>C;p.L765P) et sélectionné dans le but de possiblement expliquer le diabète du patient (Tableau III). En effet, alors que ce symptôme est non caractéristique d'une HSP7, *ARHGAP4* encode une protéine activant des GTPases Rho et fortement associée par inférence au diabète insipide (179). Avec une prédiction CADD appréciable (score = 24.5) et une AF nulle accompagnée d'une faible tolérance aux mutations (pLI = 0.98) (106), le gène représente un excellent candidat moléculaire pour expliquer ce type de pathologie. En plus de mettre de l'avant la puissance du séquençage génomique pour l'obtention d'un portrait fidèle des anomalies génétiques d'un patient, cette découverte fortuite souligne la possibilité que de multiples mutations distinctes soient à l'origine d'une partie de l'hétérogénéité clinique observée chez les maladies mendéliennes. L'application consistante de ce concept à la recherche en génomique clinique pourrait permettre de mieux définir la présentation phénotypique caractéristique de nombreuses maladies génétiques (49).

4.3.6 Expansion *ATXN2* avec possible modulateur génétique dans *ZFYVE26*

Suite à l'analyse des données WGS le meilleur candidat identifié s'est avéré être un SNV non-sens dans le gène *ZFYVE26* (c.3022G>A;p.R1008X) qui comme pour le patient MT-0012, possède un énorme potentiel pathogénique selon CADD (score = 42) et une fréquence allélique nulle (106). De plus, l'introduction du codon stop se fait dans l'exon 17, présent dans tous les isoformes majeurs de la protéine, qui a une expression ubiquitaire et très élevée dans le cervelet (132). *ZFYVE26* encode une protéine liant le phosphatidylinositol 3-phosphate, requise pour plusieurs fonctions dont la maturation de l'autophagosome où les défauts lysosomaux et d'autophagie induisent des phénotypes spastiques (180). Plus précisément, le gène est associé à HSP15, une forme AR et progressive de la maladie où plusieurs phénotypes typiques tels que le nystagmus, une faiblesse musculaire des membres inférieurs, une ataxie cérébelleuse ainsi qu'une dysarthrie concorde bien avec le patient (Tableau III), qui démontre toutefois une présentation clinique plus légère et tardive (181). L'hypothèse d'une forme AD causée par la troncation de plus de la moitié de la protéine pourrait toutefois être à l'origine d'une pénétrance variable des phénotypes (176), phénomène qui n'est pas rare pour les désordres de types HSP (53). Néanmoins, l'évaluation visuelle de l'alignement transcriptomique de *ZFYVE26* permet d'observer une faible proportion de jonctions exoniques non canoniques étant absentes des autres échantillons, et rMATS supporte faiblement un saut d'exon anormal chez MT-0013 (score = 0.146). Malgré l'absence de variants identifiables dans la région génomique adjacente à l'événement d'épissage, expliquant le non-appel du gène par SpliceAI, la nature AR de HSP15 a mené à une validation parallèle des deux événements potentiellement pathogéniques.

Alors que le SNV non-sens a été confirmé (Figure 11B), ce n'est pas le cas du possible AS, où le transcriptome de MT-0013 pour la région d'intérêt est quasi-identique au contrôle (Figure 11A). Néanmoins, une diminution de l'expression de près de 50% suggère que l'allèle contenant la mutation délétère est dégradé par NMD (Figure 11C). L'alignement RNA-seq suggère également que celle-ci est moins exprimée (8/29 lectures). Dans le cas présent, l'absence d'un second variant causant un hétérozygote composé n'est pas essentielle au diagnostic moléculaire. L'hypothèse d'une troncation hétérozygote avec effet modulateur sur le candidat *ATXN2* pourrait toutefois expliquer les phénotypes atypiques tels que la faiblesse musculaire observée chez MT-0013 (49).

L'outil EH est le seul ayant prédit 32 répétitions du STR-CAG dans l'exon 1 de *ATXN2*. Cette expansion poly-Q est une cause bien établie de SCA progressive de type 2 dont le seuil pathogénique est précisément défini : alors qu'un porteur de 31 répétitions et moins est normalement sain, un seul allèle de 33 répétitions ou plus est nécessaire pour le développement de la pathologie AD (33). Cela signifie qu'un patient porteur CAG_{32} devrait présenter une pénétrance phénotypique variable, mais relativement légère comparativement ayant une forme sévère de SCA2 (182). Ce concept est fréquemment observé dans les pathologies à expansion nucléotidiques, où un plus grand nombre de STR mène à une apparition plus précoce et sévère de la maladie, ce qui est notamment visible dans le phénomène d'anticipation héréditaire (31). Cela étant dit, outre une faiblesse musculaire, la présentation clinique de MT-0013 est typique d'une SCA2 (Tableau III). Malgré le fait que l'expansion soit contenue dans une séquence codante, le séquençage Sanger peine généralement à bien définir les régions avec STR (183), ce qui explique la vérification préliminaire restreinte à l'ADNg. Cette approche était également appuyée par le fait qu'HaplotypeCaller ainsi que STRetch n'ont pas fait l'appel de cette expansion, suggérant un possible faux positif d'EH. Toutefois, les résultats Sanger de MT-0013 confirment clairement des allèles de 32 et 31 répétitions (Figure 11C), visibles grâce à l'unique trinucléotide CAA qui est décalé d'une seule position dans les deux sens. Il est probable qu'une simple confirmation de l'expansion CAG d'*ATXN2* par un laboratoire clinique suffise au diagnostic moléculaire du patient, puisque les mécanismes pathogéniques sont déjà largement étudiés (33). Cependant, il serait pertinent d'investiguer le SNV causant p.R1008X de façon plus extensive afin de confirmer s'il y a un effet modulateur sur les phénotypes de MT-0013 (12).

4.3.7 Double SNV faux-sens avec déséquilibre allélique de *CACNA1H*

Similairement à MT-0011, peu de candidats ont été identifiés par l'analyse des données NGS du patient MT-0014. Deux SNV faux-sens ont été détecté dans le gène *CACNA1H* (c.4772G>A;p.R1591Q + c.2354A>T;p.K785M), encodant la sous-unité fonctionnelle du canal calcique Cav3.2. L'intérêt majeur pour ce candidat provient du lien indirect avec l'EA de type 2, forme de la maladie due aux variants affectant Cav2.1 (35). Contrairement à son paralogue cependant, les niveaux cérébelleux de *CACNA1H* sont beaucoup plus faibles, ayant plutôt un patron d'expression ubiquitaire qui atteint son maximum dans les tissus musculaires et

hormonaux (132, 135). Il n'est donc pas surprenant que le gène soit associé à un hyperaldostéronisme familial (184). Même si la présentation clinique ne concorde pas du tout avec le patient, l'implication de *CACNA1H* dans une forme précoce d'épilepsie est intéressante puisqu'elle démontre le potentiel de phénotypes neurologiques engendrés par des mutations faux-sens dans la séquence codante (185). Quoiqu'il en soit, le potentiel causatif de *CACNA1H* est relativement faible dû à la tolérance prédite du gène contre les mutations (pLI = 0), et à la classification « bégnine » du variant p.K785M (12). La prédiction CADD est bonne (score = 24.2), mais il y a un grand nombre de porteurs ($3.53x10^{-3}$), incluant trois homozygotes sains (106). Le second variant, avec une prédiction CADD similaire (score = 27.5), est plus intéressant malgré une AF élevée ($3.37x10^{-4}$) puisqu'aucun homozygote n'est recensé.

Les deux mutations ont été confirmées par Sanger (Figure 12B), mais c'est principalement le déséquilibre allélique observé dans l'ADNc pour c.4772G>A qui suscite un intérêt de validation supplémentaire. Puisqu'une forme AD est plus qu'improbable avec *CACNA1H*, les deux SNV doivent être confirmés afin de conserver le statut de candidat chez MT-0014, mais c.2354A>T est déjà considéré bénin selon les critères de l'ACMG (12). Cependant, il semble que seul l'allèle portant le second variant est exprimé chez le patient, agissant alors comme une mutation hémizygote (186). Dans cette optique, une validation de l'impact fonctionnel de p.R1591Q pourrait permettre une association phénotype-génotype, mais la recherche de candidats additionnels est en cours chez MT-0014.

4.4 Candidats finaux pour les deux patients du trio familial

4.4.1 SNV faux-sens pathogénique dans *SPAST*

Telle que mentionnée précédemment, l'analyse des résultats de HaplotypeCaller pour le WGS de la jeune fille a rapidement mené à l'identification d'un SNV faux-sens pathogénique présent dans tous les isoformes de *SPAST* (c.1496G>A;p.R499H) (12). En effet, le gène encode une protéine ATPase modulant la dynamique des microtubules, et la perte de l'arginine 499 affecte le domaine catalytique accomplissant cette fonction essentielle (187, 188). L'expression de *SPAST* étant ubiquitaire, quoique maximale dans le cerveau et les tissus musculaires (132), les mutations

délétères sont causatifs d'une forme AD progressive de HSP de type 4 (189). Plus spécifiquement, la mutation de l'arginine 499 cause une forme sévère et précoce qui concorde parfaitement avec la présentation clinique de la fille (190). Malgré le fait que le changement d'acide aminé le plus fréquent est R499C, plusieurs porteurs R499H ont été rapportés comme atteint de HSP4 dans la base de données ClinVar (43). Bien évidemment, la prédiction CADD est élevée (score = 34) et la fréquence allélique est nulle sur gnomAD (106). La patiente a par la suite obtenu un diagnostic moléculaire de HSP4 grâce à la confirmation subséquente d'un laboratoire clinique.

L'apparition *de novo* de la mutation *SPAST* n'est pas surprenante, puisque le père avait reçu un test génétique négatif pour ce gène plusieurs années auparavant. Cela signifie non seulement que l'hypothèse d'une cause commune avec anticipation est improbable, puisque la jeune fille présente dans tous les cas une pathologie distincte, mais aussi qu'elle a vraisemblablement compliqué l'obtention d'un diagnostic. En effet, considérant le fait que la fille présentait de nombreux phénotypes typiques d'une HSP, un panel incluant *SPAST* aurait probablement dû être effectué comme test génétique initial. Malheureusement, de nombreux gènes candidats ont plutôt été exclus par extrapolation des résultats négatifs du père. Il reste à déterminer si le duo partage tout de même la pathologie à caractère tardif étant déjà apparue chez le père.

4.4.2 Expansion *RFC1*

Plusieurs variants intéressants ont été identifiés par le pipeline d'analyse génomique chez le patient français qui présente une ataxie cérébelleuse tardive et lentement progressive. Un candidat est particulièrement intéressant dû à son association récente au désordre génétique CANVAS, forme AR de plus en plus fréquente d'ataxie tardive, surtout chez les patients de descendance européenne (145, 191). Les dysfonctions cérébelleuses et vestibulaires sont dues à une expansion bi-allélique du pentanucléotide AAAAG dans le second intron de *RFC1*. Alors que le génome consensus ne contient que 12 répétitions du STR, la majorité des patients diagnostiqués sont porteurs de plus de 400 répétitions, ayant également parfois un STR plus riche en guanine tel que AAGGG (149). La fonction du facteur de réplication est reliée à celle du gène *PCNA*, autre cause connue d'ataxie (192), mais *RFC1* participe également à la régulation de la transcription et de la réplication ADN (193). Sans surprise, son patron d'expression est donc

ubiquitaire, incluant des niveaux élevés dans le cerveau (132). Seul EH a été en mesure de détecter une expansion *RFC1* dans les données WGS du patient, avec une prédiction d'allèles de 37 et 43 répétitions. Quoique cela est supérieur au consensus du génome hg19, ces nombres sont loin d'atteindre le seuil diagnostic récemment établi (145). Il est cependant important de se rappeler qu'EH tend à sous-estimer le nombre de répétitions dans un échantillon avec expansion, ce qui est logique puisqu'aucune des courtes lectures du WGS ne couvre réellement la région STR, et l'outil doit utiliser des données extérieures pour effectuer sa prédiction.

Lors de la visualisation de l'alignement avec IGV (Figure 14A), la diminution de couverture proportionnelle au nombre d'allèles portant l'expansion appuie la prédiction d'EH. En effet, puisque les lectures ont une longueur maximale de 150pb, elles ne peuvent s'aligner à leur locus d'origine sans séquençage des régions flanquant l'expansion. Plus précisément, comme le STR *RFC1* est de 5pb, il est peu probable qu'une expansion au-delà de 25 répétitions (125pb) soit en mesure de s'aligner au locus d'intérêt, et la majorité des lectures composées à 100% du STR sont simplement perdues. C'est là que la fonction de quantification STR de STRetch est pertinente, elle qui peut tenir compte de ces lectures autrement perdues, et détecter la proportion anormalement élevée de pentanucléotides AAAAG (119). Le recrutement de plusieurs membres de la famille (Figure 14B) a permis de confirmer rapidement que l'expansion *RFC1* du père possède un potentiel pathogénique, puisque la simple amplification de la région d'intérêt par PCR suggère que le patient porte le plus grand nombre de répétitions (Figure 14C). Malgré le fait que cette méthode soit imprécise pour quantifier le nombre de STR exact, elle permet tout de même de voir que la taille de l'expansion est au minimum supérieur à 100 répétitions (500pb), et que le père semble avoir un fragment *RFC1* légèrement supérieur à ses parents ainsi que son frère. Cela est particulièrement intéressant puisque malgré l'absence d'un diagnostic pathologique chez ceux-ci, ils présentent tous de même quelques signes de capacité cérébelleuse diminuée. En effet, des phénotypes tels qu'une dysphagie tardive chez le grand-père, et des troubles de l'équilibre léger chez plusieurs membres de la famille suggèrent une atteinte pathologique partielle. Ainsi, la validation du nombre de répétitions chez ceux-ci permettrait non seulement d'obtenir un diagnostic moléculaire pour le père, mais possiblement de mieux définir le seuil pathogénique de l'expansion *RFC1*. À cet effet, le LRS offre un énorme potentiel de quantification précise (194).

5 - Conclusion et perspectives

Le but principal du projet était de développer une approche efficace de diagnostic de mutations pathogéniques chez des cas complexes par intégration de données génomiques et transcriptomiques dans un pipeline d'analyse optimisé, rapide, et facile à utiliser. Outre l'appel de variant via HaplotypeCaller sur les données RNA-seq, qui a dû être remplacé par Vardict pour une meilleure précision, la mise en place des nombreux outils s'est bien déroulée. Cela inclue l'ajout de SpliceAI, EH et STRetch, dont les capacités prédictives des outils prédécesseurs étaient généralement inefficaces malgré le grand potentiel pathogénique des AS et expansion nucléotidiques. En effet, ExpansionHunter est le seul outil ayant permis l'identification des meilleurs candidats pour MT-0013 ainsi que pour le père du trio WGS, soient *ATXN2* et *RFC1* respectivement. Dans le cas de SpliceAI, ses appels ont consolidé rapidement les candidats préliminaires *ELOVL4* et *PMPCB* à partir des données génomiques, en plus de fournir davantage d'informations sur les événements d'épissages prédits. Avec de telles performances (Figure 5; Tableau VI) (116, 124, 125), il est fort probable que ces outils s'établissent en tant que standard de la génétique clinique dans un futur rapproché.

La combinaison des informations fonctionnelles du RNA-seq au WGS s'est également avérée considérablement bénéfique, particulièrement chez MT-0009. Alors que le séquençage transcriptomique seul n'aurait pas permis l'appel du variant intronique *ELOVL4* dû à sa faible couverture dans l'ARNm, l'observation des jonctions exoniques non canoniques du gène a exacerbé le potentiel pathogénique de la mutation. Alors que la limitation majeure du WGS est l'identification d'un très grand nombre de VUS, la capacité d'observer les jonctions d'un transcrit, la présence d'expression différentielle, ou encore un déséquilibre allélique permet de classifier plus précisément l'intérêt pour chaque candidat dans un patient (70). Dans l'optique plus large du NGS, l'avantage clair de l'approche est la capacité d'analyser parallèlement de nombreuses mutations, qu'elles soient présentes dans des gènes associés aux désordres soupçonnés ou non. C'est ce qui donne lieu à l'identification de gènes candidats ayant été peu étudiés malgré leur proximité moléculaire à la pathologie d'intérêt, tels qu'*ATXN7L1* et *PMPCB*.

L'approche offre également un potentiel unique pour l'explication d'une partie de l'hétérogénéité clinique fréquemment observée dans les maladies neurologiques. En effet, la détection de modulateur génétique ou encore de mutations pathogéniques de désordres distincts représente un avantage moderne pour mieux distinguer les phénotypes caractéristiques d'une cause génétique de ceux étant atypiques (49). Alors que la faiblesse musculaire et paralysie sporadique des membres inférieures de MT-0013 pourrait être rapportée comme symptôme inhabituel d'une SCA2 avec seulement 32 répétitions CAG dans *ATXN2*, il est bien plus logique de démontrer un rôle modulateur de la troncation *ZFYVE26*, induisant vraisemblablement des phénotypes partiels de HSP15 (181, 182). Similairement, le diabète de MT-0012 n'est pas caractéristique des HSP7, cause vraisemblable de la pathologie ataxique du patient (178). Nonobstant le besoin d'une validation expérimentale supplémentaire du variant *ARHGAP4*, il serait plus pertinent de compléter cette nouvelle association plutôt que d'ajouter un phénotype atypique au diagnostic de paraplégie spastique (179). En fait, la pertinence de ce type de découverte dépasse le cadre de la génomique clinique, pouvant même bénéficier la recherche fondamentale au travers des nouvelles associations moléculaires.

En incluant le trio familial, les données de dix patients ont été traitées avec le pipeline d'analyse mis en place pour ce projet pilote. Sur ce nombre, quatre ont reçu ou sont en voie de recevoir un diagnostic moléculaire (40%), quatre sont porteurs d'excellents candidats suscitant une validation expérimentale plus poussée (40%), et deux nécessite une recherche additionnelle de cause génétique parallèlement à l'évaluation d'un candidat à potentiel faible ou modéré (20%). Quoique le nombre de participants à l'étude soit faible, le taux de succès se compare aux meilleures études concernant les ataxies (127), qui ne se limitent habituellement pas aux cas complexes, et pourrait s'améliorer à la suite des validations futures. Ces perspectives de recherche incluent notamment la modélisation des mutations d'*ATXN7L1*, *PMPCB*, *GABRP*, et *CACNA1H* dans des neurones transdifférenciés ou encore un organisme animal tel que le poisson-zèbre afin de permettre une meilleure interprétation de leur impact fonctionnel (156, 195). Les hauts niveaux d'expression *ZFYVE26* pourraient quant à eux permettre une évaluation de l'effet protéique de la troncation directement à partir des PBMC, simplifiant l'association éventuelle du variant aux phénotypes HSP15 (12). Dans le cas de *RFC1*, cause maintenant bien établie d'ARCA tardive (145), la

quantification de la nature et du nombre exact de répétitions de pentanucléotide AARRG est prioritaire au diagnostic du patient (149). La disponibilité d'échantillons provenant d'un large pedigree familial pourrait subséquemment permettre de redéfinir le seuil pathogénique de l'expansion bi-allélique intronique. À cet effet, le LRS est une option simple, rapide et précise de quantification STR dont le potentiel clinique suscite énormément d'intérêt (62, 194). Cela étant dit, la possibilité d'une investigation additionnelle du mécanisme pathogénique de l'expansion *RFC1* n'est pas écartée des perspectives futures.

Un autre point soulevé par les résultats du projet est l'importance d'une approche non biaisée lors de la recherche d'un diagnostic génétique, ce qui n'est pas toujours possible en milieu hospitalier où les ressources sont parfois limitées et la pression décisionnelle est mise sur les épaules du clinicien. Chez les deux patients du trio familial, la corrélation partielle des phénotypes neurologiques progressifs a mené à une hypothèse stricte d'expansion héréditaire avec phénomène d'anticipation. Dans l'optique où cette dernière est logique, et où de nombreuses ressources ont déjà été consumées pour le diagnostic moléculaire du père, l'hypothèse permet d'accentuer l'effort sur de nouveaux tests pour la jeune fille. Toutefois, dans un contexte où les deux parents sont sains, la présentation clinique aurait rapidement mené à un panel génétique pour les HSP qui aurait permis la détection du variant *SPAST* (189). En effet, ce gène est inclus dans l'un des premiers panels génétiques pour lequel le père a obtenu des résultats négatifs. Quoique cette découverte affecte peu la prise en charge du patient, l'obtention d'un diagnostic officiel s'accompagne de plusieurs avantages sociaux dans ce cas, bénéficiant directement la famille.

Les technologies de séquençage NGS permettent d'éviter ce type de diagnostic tardif, puisqu'il n'y a pas de contrainte directe des gènes testés (83). Bien qu'initialement plus dispendieux, l'obtention simplifiée de résultats positifs pour une plus grande proportion de patients réduit en partie les coûts associés à l'utilisation du NGS en milieu clinique. De plus, l'optimisation des méthodes expérimentales, par exemple via multiplexage, mène également à une diminution constante du prix de séquençage de chaque échantillon (196). Les résultats obtenus lors de ce projet pilote démontrent donc la pertinence de l'approche, du moins chez les patients complexes ou bien ayant une présentation clinique atypique.

D’un point de vue logistique, il n’est pas inconcevable que cette approche puisse être réalisée en milieu hospitalier. Étant à première vue plus complexe, vu la nécessité d’une manipulation rapide et particulière permettant la récolte de l’ARN en plus de l’ADN, la mise en place d’un protocole décisionnel permettrait de limiter les ressources supplémentaires nécessaire à cette approche. En effet, le traitement additionnel des échantillons sanguin pourrait être réservés aux patients où la pathologie génétique soupçonnée est relativement hétérogène, comme c’était le cas avec les EA. De plus, il serait possible d’initialement effectuer des tests traditionnels par panels, puis de procéder seulement au WGS, afin d’identifier des variants causatifs évidents tels que ceux observés chez MT-0012. L’ARN serait conservé à -80°C jusqu’à ce que les résultats préliminaires suggèrent l’intérêt de performer le RNA-seq pour son information fonctionnelle. Ce type d’optimisation de l’approche permettrait potentiellement d’améliorer les rendements de diagnostics moléculaires ainsi que les connaissances générales en génomique, tout en réduisant éventuellement le fardeau économique que représentent ces maladies rares à caractère génétique parfois complexe.

Finalement, quoique la trop faible quantité de contrôles disponibles a ralenti la méta-analyse de l’expression différentielle RNA-seq, les données générées permettront possiblement l’identification de régulateurs et des voies moléculaires communément affectées chez les patients. En effet, ce type d’analyse met généralement en évidence plusieurs gènes n’ayant pas été précédemment impliqués dans la pathogenèse ataxique, suggérant un potentiel en tant que nouvelles causes génétiques ou encore comme cibles thérapeutiques futures (197).

Références bibliographiques

1. Ferreira CR. The burden of rare diseases. Am J Med Genet A. 2019;179(6):885-92.
2. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28(2):165-73.
3. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016;48(2):126-33.
4. Tan H. Somatic mutation in noncoding regions: The sound of silence. EBioMedicine. 2020;61:103084.
5. Hashim FA, Mabrouk MS, Al-Atabany W. Review of Different Sequence Motif Finding Algorithms. Avicenna J Med Biotechnol. 2019;11(2):130-48.
6. Hoogmartens J, Hens E, Engelborghs S, Vandenberghe R, De Deyn PP, Cacace R, et al. Contribution of homozygous and compound heterozygous missense mutations in VWA2 to Alzheimer's disease. Neurobiol Aging. 2021;99:100 e17- e23.
7. Vihinen M. Systematics for types and effects of DNA variations. BMC Genomics. 2018;19(1):974.
8. Zoghbi HY, Beaudet AL. Epigenetics and Human Disease. Cold Spring Harb Perspect Biol. 2016;8(2):a019497.
9. Pajusalu S, Kahre T, Roomere H, Murumets U, Roht L, Simenson K, et al. Large gene panel sequencing in clinical diagnostics-results from 501 consecutive cases. Clin Genet. 2018;93(1):78-83.
10. Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. NPJ Genom Med. 2020;5:47.
11. Siddique N, Siddique T. Amyotrophic Lateral Sclerosis Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. GeneReviews((R)). Seattle (WA)1993.
12. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.
13. Sanford EF, Clark MM, Farnaes L, Williams MR, Perry JC, Ingulli EG, et al. Rapid Whole Genome Sequencing Has Clinical Utility in Children in the PICU. Pediatr Crit Care Med. 2019;20(11):1007-20.
14. D'Amours G, Gauthier J, Hamdan F, Meleu A, Maftei C, Soucy J-F, et al. First tier rapid whole genome sequencing increases diagnostic yield and changes management in children admitted in intensive care. 2021;132:S102.
15. Samuels ME, Orr A, Guernsey DL, Dooley K, Riddell C, Hodgkinson K, et al. Is gene discovery research or diagnosis? Genet Med. 2008;10(6):385-90.
16. Bird TD. Hereditary Ataxia Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. GeneReviews((R)). Seattle (WA)1993.

17. Koziol LF, Budding D, Andreasen N, D'Arrigo S, Bulgheroni S, Imamizu H, et al. Consensus paper: the cerebellum's role in movement and cognition. Cerebellum. 2014;13(1):151-77.
18. Schweighofer N, Doya K, Kuroda S. Cerebellar aminergic neuromodulation: towards a functional understanding. Brain Res Brain Res Rev. 2004;44(2-3):103-16.
19. Yang Y, Lisberger SG. Purkinje-cell plasticity and cerebellar motor learning are graded by complex-spike duration. Nature. 2014;510(7506):529-32.
20. Kano M, Watanabe T. Type-1 metabotropic glutamate receptor signaling in cerebellar Purkinje cells in health and disease. F1000Res. 2017;6:416.
21. Schmahmann JD. Disorders of the cerebellum: ataxia, dysmetria of thought, and the cerebellar cognitive affective syndrome. J Neuropsychiatry Clin Neurosci. 2004;16(3):367-78.
22. Jensen MB, St Louis EK. Management of acute cerebellar stroke. Arch Neurol. 2005;62(4):537-44.
23. Nussinovitch M, Prais D, Volovitz B, Shapiro R, Amir J. Post-infectious acute cerebellar ataxia in children. Clin Pediatr (Phila). 2003;42(7):581-4.
24. Chester CS, Reznick BR. Ataxia after severe head injury: the pathological substrate. Ann Neurol. 1987;22(1):77-9.
25. Ouahchi K, Arita M, Kayden H, Hentati F, Ben Hamida M, Sokol R, et al. Ataxia with isolated vitamin E deficiency is caused by mutations in the alpha-tocopherol transfer protein. Nat Genet. 1995;9(2):141-5.
26. Bidichandani SI, Delatycki MB. Friedreich Ataxia. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. GeneReviews((R)). Seattle (WA)1993.
27. Beaudin M, Klein CJ, Rouleau GA, Dupre N. Systematic review of autosomal recessive ataxias and proposal for a classification. Cerebellum Ataxias. 2017;4:3.
28. Beaudin M, Matilla-Duenas A, Soong BW, Pedroso JL, Barsottini OG, Mitoma H, et al. The Classification of Autosomal Recessive Cerebellar Ataxias: a Consensus Statement from the Society for Research on the Cerebellum and Ataxias Task Force. Cerebellum. 2019;18(6):1098-125.
29. De Braekeleer M, Giasson F, Mathieu J, Roy M, Bouchard JP, Morgan K. Genetic epidemiology of autosomal recessive spastic ataxia of Charlevoix-Saguenay in northeastern Quebec. Genet Epidemiol. 1993;10(1):17-25.
30. Klar J, Ali Z, Farooq M, Khan K, Wikstrom J, Iqbal M, et al. A missense variant in ITPR1 provides evidence for autosomal recessive SCA29 with asymptomatic cerebellar hypoplasia in carriers. Eur J Hum Genet. 2017;25(7):848-53.
31. Paulson H. Repeat expansion diseases. Handb Clin Neurol. 2018;147:105-23.
32. Wallace SE, Bean LJH. Genetic Disorders Caused by Nucleotide Repeat Expansions and Contractions. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. GeneReviews((R)). Seattle (WA)2017.
33. Pulst SM. The complex structure of ATXN2 genetic variation. Neurol Genet. 2018;4(6):e299.
34. Nakamura Y, Tagawa K, Oka T, Sasabe T, Ito H, Shiwaku H, et al. Ataxin-7 associates with microtubules and stabilizes the cytoskeletal network. Hum Mol Genet. 2012;21(5):1099-110.
35. Mantuano E, Romano S, Veneziano L, Gellera C, Castellotti B, Caimi S, et al. Identification of novel and recurrent CACNA1A gene mutations in fifteen patients with episodic ataxia type 2. J Neurol Sci. 2010;291(1-2):30-6.

36. Damaj L, Lupien-Meilleur A, Lortie A, Riou E, Ospina LH, Gagnon L, et al. CACNA1A haploinsufficiency causes cognitive impairment, autism and epileptic encephalopathy with mild cerebellar symptoms. Eur J Hum Genet. 2015;23(11):1505-12.
37. Choi KD, Choi JH. Episodic Ataxias: Clinical and Genetic Features. J Mov Disord. 2016;9(3):129-35.
38. Jen JC, Graves TD, Hess EJ, Hanna MG, Griggs RC, Baloh RW, et al. Primary episodic ataxias: diagnosis, pathogenesis and treatment. Brain. 2007;130(Pt 10):2484-93.
39. Fazeli W, Becker K, Herkenrath P, Duchting C, Korber F, Landgraf P, et al. Dominant SCN2A Mutation Causes Familial Episodic Ataxia and Impairment of Speech Development. Neuropediatrics. 2018;49(6):379-84.
40. Salles PA, Mata IF, Brunger T, Lal D, Fernandez HH. ATP1A3-Related Disorders: An Ever-Expanding Clinical Spectrum. Front Neurol. 2021;12:637890.
41. Hyman SE. Neurotransmitters. Curr Biol. 2005;15(5):R154-8.
42. Imbrici P, D'Adamo MC, Kullmann DM, Pessia M. Episodic ataxia type 1 mutations in the KCNA1 gene impair the fast inactivation properties of the human potassium channels Kv1.4-1.1/Kvbeta1.1 and Kv1.4-1.1/Kvbeta1.2. Eur J Neurosci. 2006;24(11):3073-83.
43. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D7.
44. Catterall WA. Voltage-gated calcium channels. Cold Spring Harb Perspect Biol. 2011;3(8):a003947.
45. Escayg A, De Waard M, Lee DD, Bichet D, Wolf P, Mayer T, et al. Coding and noncoding variation of the human calcium-channel beta4-subunit gene CACNB4 in patients with idiopathic generalized epilepsy and episodic ataxia. Am J Hum Genet. 2000;66(5):1531-9.
46. Choi KD, Jen JC, Choi SY, Shin JH, Kim HS, Kim HJ, et al. Late-onset episodic ataxia associated with SLC1A3 mutation. J Hum Genet. 2017;62(3):443-6.
47. de Vries B, Mamsa H, Stam AH, Wan J, Bakker SL, Vanmolkot KR, et al. Episodic ataxia associated with EAAT1 mutation C186S affecting glutamate reuptake. Arch Neurol. 2009;66(1):97-101.
48. Conroy J, McGettigan P, Murphy R, Webb D, Murphy SM, McCoy B, et al. A novel locus for episodic ataxia:UBR4 the likely candidate. Eur J Hum Genet. 2014;22(4):505-10.
49. Rahit K, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. Genes (Basel). 2020;11(3).
50. Kim JM, Ryu WS, Hwang YH, Kim JS. Aggravation of ataxia due to acetazolamide induced hyperammonaemia in episodic ataxia. J Neurol Neurosurg Psychiatry. 2007;78(7):771-2.
51. Sarva H, Shanker VL. Treatment Options in Degenerative Cerebellar Ataxia: A Systematic Review. Mov Disord Clin Pract. 2014;1(4):291-8.
52. Miyai I, Ito M, Hattori N, Mihara M, Hatakenaka M, Yagura H, et al. Cerebellar ataxia rehabilitation trial in degenerative cerebellar diseases. Neurorehabil Neural Repair. 2012;26(5):515-22.
53. Hedera P. Hereditary Spastic Paraplegia Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. GeneReviews((R)). Seattle (WA)1993.

54. Bhattacharjee S, Beauchamp N, Murray BE, Lynch T. Case series of autosomal recessive hereditary spastic paraparesis with novel mutation in SPG 7 gene. Neurosciences (Riyadh). 2017;22(4):303-7.
55. Chinnery PF. Primary Mitochondrial Disorders Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. GeneReviews((R)). Seattle (WA)1993.
56. Cabral-Costa JV, Kowaltowski AJ. Neurological disorders and mitochondria. Mol Aspects Med. 2020;71:100826.
57. Bargiela D, Shanmugarajah P, Lo C, Blakely EL, Taylor RW, Horvath R, et al. Mitochondrial pathology in progressive cerebellar ataxia. Cerebellum Ataxias. 2015;2:16.
58. Adibhatla RM, Hatcher JF. Altered lipid metabolism in brain injury and disorders. Subcell Biochem. 2008;49:241-68.
59. Klouwer FC, Berendse K, Ferdinandusse S, Wanders RJ, Engelen M, Poll-The BT. Zellweger spectrum disorders: clinical overview and management approach. Orphanet J Rare Dis. 2015;10:151.
60. Jansen GA, Waterham HR, Wanders RJ. Molecular basis of Refsum disease: sequence variations in phytanoyl-CoA hydroxylase (PHYH) and the PTS2 receptor (PEX7). Hum Mutat. 2004;23(3):209-18.
61. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018;122(1):e59.
62. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30.
63. Scacheri CA, Scacheri PC. Mutations in the noncoding genome. Curr Opin Pediatr. 2015;27(6):659-64.
64. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med. 2018;50(8):1-8.
65. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9.
66. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-48 e24.
67. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11(1):70.
68. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102-10.
69. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019;51(1):88-95.
70. Hollein A, Twardziok SO, Walter W, Hutter S, Baer C, Hernandez-Sanchez JM, et al. The combination of WGS and RNA-Seq is superior to conventional diagnostic tests in multiple myeloma: Ready for prime time? Cancer Genet. 2020;242:15-24.
71. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
72. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17(5):257-71.

73. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20(11):631-56.
74. Ross J. mRNA stability in mammalian cells. Microbiol Rev. 1995;59(3):423-50.
75. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014;13(2):397-406.
76. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 2018;59(3):253-68.
77. Lin JH, Masson E, Boulling A, Hayden M, Cooper DN, Ferec C, et al. 5' splice site GC>GT and GT>GC variants differ markedly in terms of their functionality and pathogenicity. Hum Mutat. 2020;41(8):1358-64.
78. Fresard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25(6):911-9.
79. Clark MB, Wrzesinski T, Garcia AB, Hall NAL, Kleinman JE, Hyde T, et al. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. Mol Psychiatry. 2020;25(1):37-47.
80. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008;26(10):1146-53.
81. Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genom Bioinform. 2020;2(2):lqaa037.
82. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. Dev Growth Differ. 2019;61(5):316-26.
83. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91.
84. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
85. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019;20(1):129.
86. Wick RR, Judd LM, Holt KE. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLoS Comput Biol. 2018;14(11):e1006583.
87. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
88. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9.
89. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525-7.
90. Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO. Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics. 2018;19(1):510.
91. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
92. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-15.

93. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods. 2017;14(2):135-9.
94. Musich R, Cadle-Davidson L, Osier MV. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Front Plant Sci. 2021;12:657240.
95. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.
96. Endrullat C, Glokler J, Franke P, Frohme M. Standardization and quality management in next-generation sequencing. Appl Transl Genom. 2016;10:2-9.
97. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.
98. Leger A, Leonardi T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software. 2019;4(34).
99. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
100. Tian S, Yan H, Kalmbach M, Slager SL. Impact of post-alignment processing in variant discovery from whole exome data. BMC Bioinformatics. 2016;17(1):403.
101. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.
102. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
103. Beck TF, Mullikin JC, Program NCS, Biesecker LG. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clin Chem. 2016;62(4):647-54.
104. Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu YF, McSweeney KM, et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet Med. 2015;17(10):774-81.
105. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
106. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
107. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000;28(1):352-5.
108. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7.
109. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94.
110. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4.
111. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.
112. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019;10(1):3240.

113. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333-i9.
114. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84.
115. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12(4):e1004873.
116. Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. Recent advances in the detection of repeat expansions with short-read next-generation sequencing. F1000Res. 2018;7.
117. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35(22):4754-6.
118. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt J, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020;21(1):102.
119. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. 2018;19(1):121.
120. Mehmood A, Laiho A, Venalainen MS, McGlinchey AJ, Wang N, Elo LL. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform. 2020;21(6):2052-65.
121. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111(51):E5593-601.
122. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012;22(10):2008-17.
123. Lord J, Baralle D. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. Front Genet. 2021;12:689892.
124. Ha C, Kim JW, Jang JH. Performance Evaluation of SpliceAI for the Prediction of Splicing of NF1 Variants. Genes (Basel). 2021;12(9).
125. Rowlands C, Thomas HB, Lord J, Wai HA, Arno G, Beaman G, et al. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. Sci Rep. 2021;11(1):20607.
126. Lesurf R, Persad G, Mital S. WHOLE GENOME SEQUENCING IDENTIFIES NOVEL CRYPTIC SPLICE SITE VARIANTS IN CHILDREN WITH CARDIOMYOPATHY. Canadian Journal of Cardiology. 2021;37(10, Supplement):S68-S9.
127. Fogel BL, Lee H, Deignan JL, Strom SP, Kantarci S, Wang X, et al. Exome sequencing in the clinical diagnosis of sporadic or familial cerebellar ataxia. JAMA Neurol. 2014;71(10):1237-46.
128. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA. 2014;312(18):1880-7.
129. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108.
130. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.

131. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics. 2016;54:1 30 1-1 3.
132. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. 2018;46(D1):D246-D51.
133. Wu C, Jin X, Tsueng G, Afrasiabi C, Su AI. BioGPS: building your own mash-up of gene annotations and expression profiles. Nucleic Acids Res. 2016;44(D1):D313-6.
134. Uhlen M, Karlsson MJ, Zhong W, Tebani A, Pou C, Mikes J, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science. 2019;366(6472).
135. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 2015;16:22.
136. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-D13.
137. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840-D5.
138. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-6.
139. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33(Web Server issue):W306-10.
140. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.
141. Roy JG, McElhaney JE, Verschoor CP. Reliable reference genes for the quantification of mRNA in human T-cells and PBMCs stimulated with live influenza virus. BMC Immunol. 2020;21(1):4.
142. Panahi Y, Salasar Moghaddam F, Ghasemi Z, Hadi Jafari M, Shervin Badv R, Eskandari MR, et al. Selection of Suitable Reference Genes for Analysis of Salivary Transcriptome in Non-Syndromic Autistic Male Children. Int J Mol Sci. 2016;17(10).
143. Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 2020;48(D1):D77-D83.
144. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J Biol Chem. 2020;295(13):4134-70.
145. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51(4):649-58.
146. Sirp A, Leite K, Tuvikene J, Nurm K, Sepp M, Timmusk T. The Fuchs corneal dystrophy-associated CTG repeat expansion in the TCF4 gene affects transcription from its alternative promoters. Sci Rep. 2020;10(1):18424.
147. Tazelaar GHP, Dekker AM, van Vugt J, van der Spek RA, Westeneng HJ, Kool L, et al. Association of NIPA1 repeat expansions with amyotrophic lateral sclerosis in a large international cohort. Neurobiol Aging. 2019;74:234 e9- e15.
148. Barratt S, Kendrick AH, Buchanan F, Whittle AT. Central hypoventilation with PHOX2B expansion mutation presenting in adulthood. Thorax. 2007;62(10):919-20.

149. Akcimen F, Ross JP, Bourassa CV, Liao C, Rochefort D, Gama MTD, et al. Investigation of the RFC1 Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. Front Genet. 2019;10:1219.
150. Danis D, Jacobsen JOB, Carmody LC, Gargano MA, McMurry JA, Hegde A, et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. Am J Hum Genet. 2021;108(9):1564-77.
151. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet Med. 2020;22(6):1005-14.
152. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. Nat Rev Genet. 2016;17(7):407-21.
153. Lesage S, Drouet V, Majounie E, Deramecourt V, Jacoupy M, Nicolas A, et al. Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. Am J Hum Genet. 2016;98(3):500-13.
154. Browne DL, Gancher ST, Nutt JG, Brunt ER, Smith EA, Kramer P, et al. Episodic ataxia/myokymia syndrome is associated with point mutations in the human potassium channel gene, KCNA1. Nat Genet. 1994;8(2):136-40.
155. Po S, Roberds S, Snyders DJ, Tamkun MM, Bennett PB. Heteromultimeric assembly of human potassium channels. Molecular basis of a transient outward current? Circ Res. 1993;72(6):1326-36.
156. Mollinari C, Zhao J, Lupacchini L, Garaci E, Merlo D, Pei G. Transdifferentiation: a new promise for neurodegenerative diseases. Cell Death Dis. 2018;9(8):830.
157. Stevanin G, Giunti P, Belal GD, Durr A, Ruberg M, Wood N, et al. De novo expansion of intermediate alleles in spinocerebellar ataxia 7. Hum Mol Genet. 1998;7(11):1809-13.
158. Saito K, Tautz L, Mustelin T. The lipid-binding SEC14 domain. Biochim Biophys Acta. 2007;1771(6):719-26.
159. Tracey TJ, Steyn FJ, Wolvetang EJ, Ngo ST. Neuronal Lipid Metabolism: Multiple Pathways Driving Functional Outcomes in Health and Disease. Front Mol Neurosci. 2018;11:10.
160. Anantharaman V, Aravind L. The GOLD domain, a novel protein module involved in Golgi function and secretion. Genome Biol. 2002;3(5):research0023.
161. Agbaga MP, Stiles MA, Brush RS, Sullivan MT, Machalinski A, Jones KL, et al. The Elovl4 Spinocerebellar Ataxia-34 Mutation 736T>G (p.W246G) Impairs Retinal Function in the Absence of Photoreceptor Degeneration. Mol Neurobiol. 2020;57(11):4735-53.
162. Deak F, Anderson RE, Fessler JL, Sherry DM. Novel Cellular Functions of Very Long Chain-Fatty Acids: Insight From ELOVL4 Mutations. Front Cell Neurosci. 2019;13:428.
163. Ding J, Miao QF, Zhang JW, Guo YX, Zhang YX, Zhai QX, et al. H258R mutation in KCNAB3 gene in a family with genetic epilepsy and febrile seizures plus. Brain Behav. 2020;10(12):e01859.
164. Polster A, Perni S, Bichraoui H, Beam KG. Stac adaptor proteins regulate trafficking and function of muscle and neuronal L-type Ca2+ channels. Proc Natl Acad Sci U S A. 2015;112(2):602-6.
165. Matthews E, Labrum R, Sweeney MG, Sud R, Haworth A, Chinnery PF, et al. Voltage sensor charge loss accounts for most cases of hypokalemic periodic paralysis. Neurology. 2009;72(18):1544-7.

166. Hyun YS, Park HJ, Heo SH, Yoon BR, Nam SH, Kim SB, et al. Rare variants in methionyl- and tyrosyl-tRNA synthetase genes in late-onset autosomal dominant Charcot-Marie-Tooth neuropathy. Clin Genet. 2014;86(6):592-4.
167. Bayat V, Thiffault I, Jaiswal M, Tetreault M, Donti T, Sasarman F, et al. Mutations in the mitochondrial methionyl-tRNA synthetase cause a neurodegenerative phenotype in flies and a recessive ataxia (ARSAL) in humans. PLoS Biol. 2012;10(3):e1001288.
168. Kaminska M, Shalak V, Mirande M. The appended C-domain of human methionyl-tRNA synthetase has a tRNA-sequestering function. Biochemistry. 2001;40(47):14309-16.
169. Vogtle FN, Brandl B, Larson A, Pendziwiat M, Friederich MW, White SM, et al. Mutations in PMPCB Encoding the Catalytic Subunit of the Mitochondrial Presequence Protease Cause Neurodegeneration in Early Childhood. Am J Hum Genet. 2018;102(4):557-73.
170. Choquet K, Zurita-Rendon O, La Piana R, Yang S, Dicaire MJ, Care4Rare C, et al. Autosomal recessive cerebellar ataxia caused by a homozygous mutation in PMPCA. Brain. 2016;139(Pt 3):e19.
171. Sanchez-Ferrero E, Coto E, Beetz C, Gamez J, Corao AI, Diaz M, et al. SPG7 mutational screening in spastic paraplegia patients supports a dominant effect for some mutations and a pathogenic role for p.A510V. Clin Genet. 2013;83(3):257-62.
172. Greene AW, Grenier K, Aguileta MA, Muise S, Farazifard R, Haque ME, et al. Mitochondrial processing peptidase regulates PINK1 processing, import and Parkin recruitment. EMBO Rep. 2012;13(4):378-85.
173. Sung HY, Yang SD, Ju W, Ahn JH. Aberrant epigenetic regulation of GABRP associates with aggressive phenotype of ovarian cancer. Exp Mol Med. 2017;49(5):e335.
174. Gazquez I, Lopez-Escamez JA. Genetics of recurrent vertigo and vestibular disorders. Curr Genomics. 2011;12(6):443-50.
175. Rinaldo L, Hansel C. Ataxias and cerebellar dysfunction: involvement of synaptic plasticity deficits? Funct Neurol. 2010;25(3):135-9.
176. Torella A, Zanobio M, Zeuli R, Del Vecchio Blanco F, Savarese M, Giugliano T, et al. The position of nonsense mutations can predict the phenotype severity: A survey on the DMD gene. PLoS One. 2020;15(8):e0237803.
177. Pfeffer G, Pyle A, Griffin H, Miller J, Wilson V, Turnbull L, et al. SPG7 mutations are a common cause of undiagnosed ataxia. Neurology. 2015;84(11):1174-6.
178. Choquet K, Tetreault M, Yang S, La Piana R, Dicaire MJ, Vanstone MR, et al. SPG7 mutations explain a significant proportion of French Canadian spastic ataxia cases. Eur J Hum Genet. 2016;24(7):1016-21.
179. Bai Y, Chen Y, Kong X. Contiguous 22.1-kb deletion embracing AVPR2 and ARHGAP4 genes at novel breakpoints leads to nephrogenic diabetes insipidus in a Chinese pedigree. BMC Nephrol. 2018;19(1):26.
180. Vantaggiato C, Panzeri E, Castelli M, Citterio A, Arnoldi A, Santorelli FM, et al. ZFYVE26/SPASTIZIN and SPG11/SPATACSIN mutations in hereditary spastic paraplegia types AR-SPG15 and AR-SPG11 have different effects on autophagy and endocytosis. Autophagy. 2019;15(1):34-57.
181. Kara E, Tucci A, Manzoni C, Lynch DS, Elpidorou M, Bettencourt C, et al. Genetic and phenotypic characterization of complex hereditary spastic paraplegia. Brain. 2016;139(Pt 7):1904-18.

182. Cancel G, Durr A, Didierjean O, Imbert G, Burk K, Lezin A, et al. Molecular and clinical correlations in spinocerebellar ataxia 2: a study of 32 families. Hum Mol Genet. 1997;6(5):709-15.
183. de Leeuw RH, Garnier D, Kroon R, Horlings CGC, de Meijer E, Buermans H, et al. Diagnostics of short tandem repeat expansion variants using massively parallel sequencing and componential tools. Eur J Hum Genet. 2019;27(3):400-7.
184. Scholl UI, Stolting G, Nelson-Williams C, Vichot AA, Choi M, Loring E, et al. Recurrent gain of function mutation in calcium channel CACNA1H causes early-onset hypertension with primary aldosteronism. Elife. 2015;4:e06315.
185. Heron SE, Khosravani H, Varela D, Bladen C, Williams TC, Newman MR, et al. Extended spectrum of idiopathic generalized epilepsies associated with CACNA1H functional variants. Ann Neurol. 2007;62(6):560-8.
186. Falkenberg KD, Braverman NE, Moser AB, Steinberg SJ, Klouwer FCC, Schluter A, et al. Allelic Expression Imbalance Promoting a Mutant PEX6 Allele Causes Zellweger Spectrum Disorder. Am J Hum Genet. 2017;101(6):965-76.
187. Errico A, Ballabio A, Rugarli EI. Spastin, the protein mutated in autosomal dominant hereditary spastic paraplegia, is involved in microtubule dynamics. Hum Mol Genet. 2002;11(2):153-63.
188. Evans KJ, Gomes ER, Reisenweber SM, Gundersen GG, Lauring BP. Linking axonal degeneration to microtubule remodeling by Spastin-mediated microtubule severing. J Cell Biol. 2005;168(4):599-606.
189. Meijer IA, Hand CK, Cossette P, Figlewicz DA, Rouleau GA. Spectrum of SPG4 mutations in a large collection of North American families with hereditary spastic paraplegia. Arch Neurol. 2002;59(2):281-6.
190. McDermott CJ, Burness CE, Kirby J, Cox LE, Rao DG, Hewamadduma C, et al. Clinical features of hereditary spastic paraplegia due to spastin mutation. Neurology. 2006;67(1):45-51.
191. Dominik N, Galassi Deforie V, Cortese A, Houlden H. CANVAS: a late onset ataxia due to biallelic intronic AAGGG expansions. J Neurol. 2021;268(3):1119-26.
192. Baple EL, Chambers H, Cross HE, Fawcett H, Nakazawa Y, Chioza BA, et al. Hypomorphic PCNA mutation underlies a human DNA repair disorder. J Clin Invest. 2014;124(7):3137-46.
193. Li Y, Gan S, Ren L, Yuan L, Liu J, Wang W, et al. Multifaceted regulation and functions of replication factor C family in human cancers. Am J Cancer Res. 2018;8(8):1343-55.
194. Nakamura H, Doi H, Mitsuhashi S, Miyatake S, Katoh K, Frith MC, et al. Long-read sequencing identifies the pathogenic nucleotide repeat expansion in RFC1 in a Japanese case of CANVAS. J Hum Genet. 2020;65(5):475-80.
195. Quelle-Regaldie A, Sobrido-Camean D, Barreiro-Iglesias A, Sobrido MJ, Sanchez L. Zebrafish Models of Autosomal Dominant Ataxias. Cells. 2021;10(2).
196. Aynaud MM, Hernandez JJ, Barutcu S, Braunschweig U, Chan K, Pearson JD, et al. A multiplexed, next generation sequencing platform for high-throughput detection of SARS-CoV-2. Nat Commun. 2021;12(1):1405.
197. Butterfield RJ, Dunn DM, Hu Y, Johnson K, Bonnemann CG, Weiss RB. Transcriptome profiling identifies regulators of pathogenesis in collagen VI related muscular dystrophy. PLoS One. 2017;12(12):e0189664.